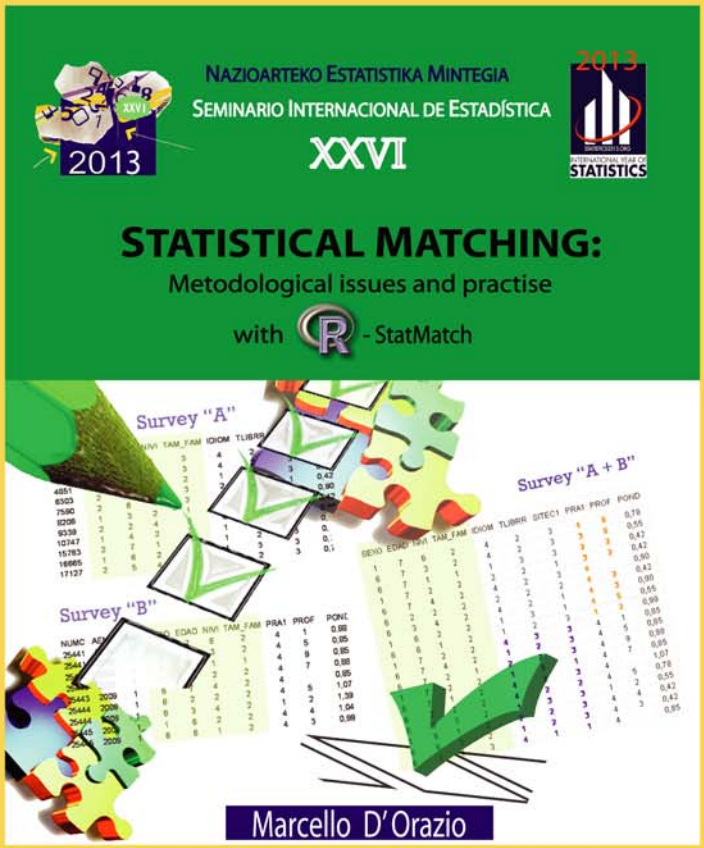


NAZIOARTEKO ESTATISTIKA MINTEGIA

SEMINARIO INTERNACIONAL DE ESTADISTICA

Statistical Matching:

Methodological Issues and Practice with R-StatMach

Marcello D´Orazio

Eustat
EUSKAL ESTATISTIKA ERAKUNDEA
INSTITUTO VASCO DE ESTADÍSTICA

55

Inkesta-lotzea:
Alderdi Metodologikoak eta R-StatMatch-ekin Praktika

Statistical matching :
Metodological issues and practice with R-StatMatch

Enlace de encuestas:
Aspectos metodológicos y práctica con R-StatMatch

Marcello D´Orazio

Department of National Accounts and Economic Statistics

Italian National Statistical Institute – Istat

E-mail: madorazi@istat.es

AURKEZPENA

Urtez urte, Nazioarteko Estatistika Mintegia antolatzean, hainbat helburu bete nahi ditu EUSTAT- Euskal Estatistika Erakundea-k:

- Unibertsitatearekiko eta, batez ere, Estatistika-Sailekiko lankidetza bultzatzea.
- Funtzionarioen, irakasleen, ikasleen eta estatistikaren alorrean interesatuta egon daitezkeen guztien lanbide-hobekuntza erraztea.
- Estatistika alorrean mundu mailan abangoardian dauden irakasle eta ikertzaile ospetsuak Euskadira ekartzea, horrek eragin ona izango baitu, zuzeneko harremanei eta esperientziak ezagutzeari dagokienez.

Aurtengoa antolatzen dugun **XXVI.** edizioa da eta guretzako mundu mailako ikertzaile aintzindari ospetsuak, estatistika alorrean, Euskadira etortzea lortu izana lorpen handia da. Hona hemen orain arte parte hartu izan duten aditu guztien zerrenda :

<<Enrique Cansado, C.R. Rao, Vic Barnett, Patrick Clapier, David J.Finney, Francisco Azorin, Jose Luis Sanchez Crespo, Gildas Roy, Jacqueline Fourastie, Richard Platek, Leslie Kish, Inmaculada Gallastegui, Marti Parellada, Luis Carlos Silva, Laurence Hausler, Felix Saltor, Adam Marton, Vivette Salvy, Graham Kalton, Wouter J. Keller/Jeike G. Bethlehem, John C. Duffy, Gad Nathan, Ildefonso Villa/ María Soledad Bravo, Estelle Bee Dagum, Camilo Dagum, Vicente Anton Valero, Leopold Granquist, Malka Kantorowich, Hans Petterson, Edwin Diday, Roger Phan Tan Luu, Karl G. Jöreskog, Antoni Espasa, Bernard Grais/Aloïs Van Bastelaer/André Persenaire, AlfonsoBarrada/Mercedes Alcalde/Cor N.Gorter, Kennet Hugh Pollock, Wolfgang Mohr/ Pekka Myrskylä, Albert Prat / Pere Grima, Roberto Escuder, David Morganstein, Johanna Varjonen, Jon N.K.Rao, Lawrence H.Cox, Jay Ver Hoef, Ray Chambers, José Miguel Bernardo, Edith de Leeuw , Peter Lynn, Eric Rancourt, Eric Schulte, William E. Yancey, Sixten Lundström/Carl-Erik Särndal, Stefano Tarantola/ Massimiliano Mascherini, Pedro M. Valero, Yves Tillé, Mick P. Couper , Christine Thomas-Agnan and this year Marcello D´ Orazio. >>

Jarduera osagarri gisa eta interesatuta egon litezkeen ahalik eta pertsona eta erakunde gehienetara iristearren, ikastaro horietako txostenak eta material guztia Eustat-eko web orrian, **www.eustat.es** eskuragarri dago**,** horrela, gai horri buruzko ezagutza zabaltzen dugularik.

Vitoria-Gasteiz, azaroak 2013

JOSU IRADI ARRIETA
EUSTAT-eko Zuzendari Nagusia

PRESENTATION

Every year, in promoting the International Statistical Seminar, EUSTAT -The Basque Statistics Institute- wishes to achieve several aims:

- Encourage the collaboration with the universities, especially with their statistical departments.
- Facilitate the professional recycling of civil servants, university teachers, students and whoever else may be interested in the statistical field.
- Bring to the Basque Country illustrious professors and investigators in the vanguard of statistical subjects, on a worldwide level, with the subsequent positive effect of encouraging direct relationships and sharing knowledge of experiences.

This year we are holding the **XXVIth** edition and it is a great achievement for us to have succeeded in bringing pioneering researchers in statistical matters, recognised worldwide to the Basque Country.
We have a list of celebrated individuals who have participated up to this point:

<<Enrique Cansado, C.R. Rao, Vic Barnett, Patrick Clapier, David J.Finney, Francisco Azorin, Jose Luis Sanchez Crespo, Gildas Roy, Jacqueline Fourastie, Richard Platek, Leslie Kish, Inmaculada Gallastegui, Marti Parellada, Luis Carlos Silva, Laurence Hausler, Felix Saltor, Adam Marton, Vivette Salvy, Graham Kalton, Wouter J. Keller/Jeike G. Bethlehem, John C. Duffy, Gad Nathan, Ildefonso Villa/ María Soledad Bravo, Estelle Bee Dagum, Camilo Dagum, Vicente Anton Valero, Leopold Granquist, Malka Kantorowich, Hans Petterson, Edwin Diday, Roger Phan Tan Luu, Karl G. Jöreskog, Antoni Espasa, Bernard Grais/Aloïs Van Bastelaer/André Persenaire, AlfonsoBarrada/Mercedes Alcalde/Cor N.Gorter, Kennet Hugh Pollock, Wolfgang Mohr/ Pekka Myrskylä, Albert Prat / Pere Grima, Roberto Escuder, David Morganstein, Johanna Varjonen, Jon N.K.Rao, Lawrence H.Cox, Jay Ver Hoef, Ray Chambers, José Miguel Bernardo, Edith de Leeuw , Peter Lynn, Eric Rancourt, Eric Schulte, William E. Yancey, Sixten Lundström/Carl-Erik Särndal, Stefano Tarantola/ Massimiliano Mascherini, Pedro M. Valero, Yves Tillé, Mick P. Couper , Christine Thomas-Agnan and this year Marcello D´ Orazio. >>

As a complementary action and in order to reach the greatest possible number of interested individuals and institutions, you have all the material and workbooks on the Eustat website **www.eustat.es**, to thus contribute to the expansion of knowledge on this subject.

Vitoria-Gasteiz, November 2013

JOSU IRADI ARRIETA
General Director of EUSTAT

PRESENTACIÓN

Año tras año, al promover los Seminarios Internacionales de Estadística, el EUSTAT-Instituto Vasco de Estadística- pretende cubrir varios objetivos:

- Fomentar la colaboración con la Universidad y en especial con los Departamentos de Estadística.
- Facilitar el reciclaje profesional de funcionarios, profesores, alumnos y cuantos puedan estar interesados en el campo estadístico.
- Traer a Euskadi a ilustres profesores e investigadores de vanguardia en materia estadística, a nivel mundial, con el consiguiente efecto positivo en cuanto a la relación directa y conocimiento de experiencias.

Este año celebramos la **XXVI[ª]** edición y para nosotros es todo un logro haber conseguido traer a Euskadi a investigadores pioneros y reconocidos en materia estadística a nivel mundial.
He aquí un listado de las personas célebres que han participado hasta ahora:

<<Enrique Cansado, C.R. Rao, Vic Barnett, Patrick Clapier, David J.Finney, Francisco Azorin, Jose Luis Sanchez Crespo, Gildas Roy, Jacqueline Fourastie, Richard Platek, Leslie Kish, Inmaculada Gallastegui, Marti Parellada, Luis Carlos Silva, Laurence Hausler, Felix Saltor, Adam Marton, Vivette Salvy, Graham Kalton, Wouter J. Keller/Jeike G. Bethlehem, John C. Duffy, Gad Nathan, Ildefonso Villa/ María Soledad Bravo, Estelle Bee Dagum, Camilo Dagum, Vicente Anton Valero, Leopold Granquist, Malka Kantorowich, Hans Petterson, Edwin Diday, Roger Phan Tan Luu, Karl G. Jöreskog, Antoni Espasa, Bernard Grais/Aloïs Van Bastelaer/André Persenaire, AlfonsoBarrada/Mercedes Alcalde/Cor N.Gorter, Kennet Hugh Pollock, Wolfgang Mohr/ Pekka Myrskylä, Albert Prat / Pere Grima, Roberto Escuder, David Morganstein, Johanna Varjonen, Jon N.K.Rao, Lawrence H.Cox, Jay Ver Hoef, Ray Chambers, José Miguel Bernardo, Edith de Leeuw , Peter Lynn, Eric Rancourt, Eric Schulte, William E. Yancey, Sixten Lundström/Carl-Erik Särndal, Stefano Tarantola/ Massimiliano Mascherini, Pedro M. Valero, Yves Tillé, Mick P. Couper , Christine Thomas-Agnan and this year Marcello D´ Orazio. >>

Como actuación complementaria y para llegar al mayor número posible de personas e Instituciones interesadas, tenéis disponibles todo el material y ejemplares de dichas formaciones en la página web de Eustat, **www.eustat.es**, para contribuir así a la expansión del conocimiento en esta materia.

Vitoria-Gasteiz, noviembre 2013

JOSU IRADI ARRIETA
Director General de EUSTAT

BIOGRAFI OHARRAK

Marcello D'Orazio Italiako Estatistika Institutu Nazionaleko ikerlari nagusietako bat da estatistikaren metodologiaren gaian. Estatistikan doktorea da Bariko Unibertsitatetik (Italia). Une honetan, nekazaritza-inkesten metodologia eta estatistika-alderaketa dira bere ikerketa-gai nagusiak. R inguruneetarako pentsatuta dagoen *Statistical Matching, Theory and Practice* izeneko monografiaren egileetako bat da.

Iraganean, inkestetako datuen kalitatea eta laginketa-teknikak ikertu izan ditu, eta aditu gisa parte hartu izan du nazioarteko hainbat lankidetza-proiektutan. R ingurunean aditua da, R-n inkesta-lotzea garatzeko «StatMatch» paketea sortu zuen, eta gaur egun bera arduratzen da pakete horren mantenimenduaz.

BIOGRAPHICAL SKETCH

Marcello D'Orazio is a Senior Research in statistical methodology at the Italian National Statistical Institute. He has a Ph.D. in statistics from University of Bari (Italy). His current research interests focus on methodologies for agriculture surveys and statistical matching; he is co-author of monograph "Statistical Matching, Theory and Practice" for the R environment.

In past, he has conducted research on quality of survey data and sampling techniques, and was involved as expert in various international cooperation projects. He is an expert of the R environment, and developed and currently maintain the package "StatMatch" for performing statistical matching in R.

NOTAS BIOGRÁFICAS

Marcello D'Orazio es un investigador principal de metodología estadística en el Instituto Nacional de Estadística Italiano. Posee un Doctorado en Estadística por la Universidad de Bari (Italia). Actualmente, sus intereses de investigación se centran en la metodología de las encuestas agrícolas y el cotejo estadístico. Es coautor de la monografía *Statistical Matching, Theory and Practice*, para el entorno R.

En el pasado desarrolló investigaciones respecto a la calidad de los datos de encuestas y técnicas de muestreo, y estuvo implicado como experto en diversos proyectos de cooperación internacional. Es un experto en el entorno R y desarrolló y mantiene en la actualidad el paquete «StatMatch» para el desarrollo del enlace de encuesta en R.

Index

1. Introduction

In recent years there has been a continuous growing demand for timely and accurate statistics. This demand cannot be met just by improving existing surveys or launching new surveys. Planning a new survey is costly and requires a nonnegligible time to plan and execute the various phases. On the other hand, if the survey form of an existing surveys is revised and enlarged, it may result a too long and complex questionnaire which, as a consequence, can determine a loss of the accuracy of the collected data. In both the cases, to improve an existing survey or to launch a new survey, there is the risk of increasing the burden on the respondents. For these reasons, and due to the growth of data made available to external users by government agencies or private companies, statisticians pointed their attention to techniques aimed at deriving the required statistics by means of *data integration* methods. In general, there two broad classes of data integration methods:

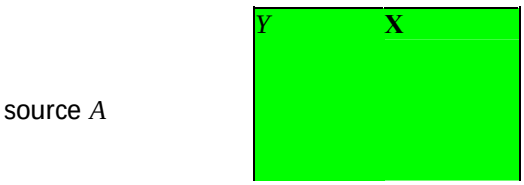
- *record linkage*
- *statistical matching*

Record linkage methods (in the wide sense of the term) aim at identifying the same units in the available data sources. The integration is based on opportune units' identifiers; if the unit identifier is a unique code (e.g. Personal Identification Number) recorded without errors then the *exact linkage* or *exact matching* can be performed. The *probabilistic record linkage* should be applied when the identifying code is affected by errors or no unique code there exists and the unit can be identified by a set of key variables (name, surname, gender, birth date, etc.). the probabilistic record linkage estimates the probability that a couple of records from different data sources refers to the same unit given the values observed on them for the key variables.

When the units lack of identifiers, the two data sources can be integrated by *statistical matching* methods. The term *statistical matching* (or *data fusion*) does not simply refer to the integration of data sources at unit level. More in general, it refers to a series of methods that use the two (or more) available data sources (usually samples), referred to the same target population, with the aim of studying the relationship among variables not jointly observed in a single data source. Usually the data sources are relatively small independent samples selected from the same target population and therefore the chance of observing the same unit in both of them is close to zero.

2. Objectives and approaches to statistical matching

In the basic statistical matching (hereafter denoted as SM) framework, there are two data sources *A* and *B* sharing a set of variables *X*, while the variable *Y* is available only in *A* and the variable *Z* is observed just in *B*. The *X* variables are common to both the data sources, while the variables *Y* and *Z* are not jointly observed.



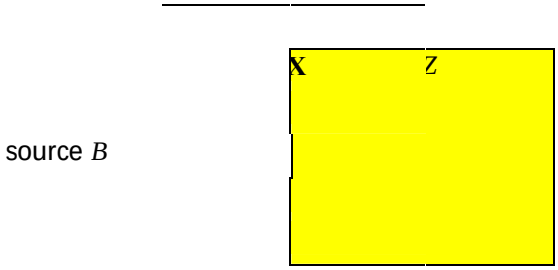


Figure 1. – Traditional Statistical Matching Framework

The objective of statistical matching consists in investigating the relationship between *Y* and *Z* at "micro" or "macro" level (D'Orazio *et al.*, 2006b).

In the *micro* case the SM aims at creating a *synthetic data source* in which all the variables, *X*, *Y* and *Z*, are available. The synthetic data set can be obtained by concatenating *A* and *B* and then filling in the missing values, i.e. *Z* in *A* and *Y* and *B*. This concatenated data set will have $n_{conc} = n_A + n_B$ units, being n_A and n_B the units in *A* and *B* respectively; therefore it is expected to provide more accurate estimates as far as the *X* variables are concerned. An alternative way of working derives the synthetic data set by considering just a data set, say *A*, and imputing in it the missing variable (*Z*) by using the information available in the other one (*B* in this case). In such example, *A* plays the role of *recipient*, while *B* is the *donor*. In this context it is important to decide which data set should be the recipient; it, once filled in with the missing variables, will be the basis of further statistical analyses; a logic choice seems that of using the larger one because it would provide more accurate results. Unfortunately, such a way of working in SM may provide inaccurate results, especially when the sizes of the two data sources are very different ($n_A >> n_B$). In practice, there is a high risk that the distribution of the imputed variable does not reflect the original one (estimated form the donor data set).

When the objective is *macro*, the data sources are matched to derive an estimate of the parameters of interest, e.g. the correlation coefficient between *Y* and *Z* or the contingency table $Y \times Z$ etc. Such an result can be achieved without necessarily passing through the integration of the data sources, e.g. by creating the synthetic data set.

A *parametric approach* to SM requires the explicit adoption of a model for the joint distribution of (*X*,*Y*,*Z*); obviously, if the model is wrongly specified the results will not be reliable. The *nonparametric approach* is more flexible in handling complex situations (mixed type variables) and it is straightforward when the objective is micro. The two approaches, parametric and nonparametric, can be mixed: first a parametric model is assumed and its parameters are estimated then a the synthetic data set is derived through a nonparametric micro approach. In this manner the advantages of both parametric and nonparametric approaches are maintained: the model is parsimonious while nonparametric techniques offer protection against model misspecification. The Table 1 provides a summary of the objectives and approaches to SM (D'Orazio *et al.*, 2008).

Table 1 – Objectives and approaches to statistical matching

Objectives of Statistical Matching	Approaches to Statistical Matching		
	Parametric	Nonparametric	Mixed
MAcro	Yes	Yes	No
MIcro	Yes	Yes	Yes

A further issue to take into account when planning the application of SM concerns the type of inference to carry out. Most of the SM techniques assume that *A* and *B* are random samples of independent and identically distributed (i.i.d.) observations selected from the same infinite population. In practice, the values assumed by *X*, *Y* and *Z* on the various units are independent outcomes of a set of random variables whose joint distribution follows a given (known or unknown) model (*model based inference*). Unfortunately in most of the cases the available data come from complex sample surveys carried out on the same finite population. In this case the set of the values assumed by *X*, *Y* and *Z* for each unit in the finite population are usually viewed as fixed values and not outcomes of random variables. The randomness is introduced by the probability criterion used to select the sample from the population. In practice, in this case the inference is based on the *sampling design* (involving stratification and/or clustering) used to select the sample from the *sampling frame*, i.e. the list which is supposed to contain all the units belonging to the target population, and for this reason inference is said *design based*.

In some circumstances it is possible to ignore the probability distribution induced by the sampling design and treat the data as coming from an i.i.d. sample (inferences are carried out in the traditional model based framework). This is possible if the sample has been selected from the sampling frame by means of simple random sampling and, more in general, if the underlying sampling design is *noninformative*, i.e. the sampling design does not depend on the target values when conditioning to the design variables (strata, etc.). Ignoring the sampling design when these conditions do not hold may lead to unreliable results (see Section 13.6 in Sarndal *et al.*, 1992).

For the sake of simplicity, in the next Sections, it will be assumed that *A* and *B* are simple random samples of i.i.d. observations referred to the same target population. The case of SM methods explicitly accounting for finite population sampling design will be addressed in the Section 6.

Section 3 will present methods to perform SM when assuming the independence of *Y* and *Z* conditioned to the common *X* variables. Section 4 will show how to bypass the conditional independence assumption by means of auxiliary information. The Section 5 will present an alternative approach to the SM based on the evaluation of the *uncertainty* due to not having observed jointly *Y* and *Z*. Section 6 will discuss SM of data arising from complex sample surveys carried on finite populations. Finally Section 7 will give some hints concerning the preliminary and final steps in a SM application.

2.1 Software to perform Statistical Matching

In past, one of the major difficulties in applying SM techniques was the lack of software packages implementing such techniques. When the objective of SM is micro, given the resemblance with an imputation problem, often the synthetic data has been derived by using/adapting software packages developed for the imputation of missing values. Sometimes, in more complex situations, when the interest is on estimating a parameter or the uncertainty concerning it, *ad hoc* software code has been provided jointly with research papers or books. Unfortunately, in most of the cases such code can be used just in very specific situations (fixed number of variables, etc.).

For these reasons it was decided to develop **StatMatch** (D'Orazio, 2012), a specific package that implements most of the commonly used SM techniques. StatMatch does not

come as stand-alone software but it is an open source additional package for the R environment (R Core Team, 2013). The first version of **StatMatch** released on the repositories of the Comprehensive R Archive Network (CRAN) dates back to 2008. This version was based on R codes provided in the Appendix of D'Orazio *et al.* (2006b). Since then a number of updates have been released. A valuable contribution to the improvement of **StatMatch** is the work done within the Eurostat's ESSnet project on "Data Integration" (D'Orazio, 2011b).

In the next Sections most of the methods presented will be followed by examples of R code calling the functions made available by **StatMatch** and others R packages helpful in the application of SM methods.

3. Statistical matching under the conditional independence assumption

In the traditional SM framework when only *A* and *B* are available, all the SM methods (parametric, nonparametric and mixed) that use the set of common variables *X* to match *A* and *B*, implicitly assume the *conditional independence* (CI) of *Y* and *Z* given *X*. In particular the CI assumption implies that the joint probability density function for *X*, *Y* and *Z* can be factorized in the following manner:

$$f(x,y,z) = f(y,z|x)\, f(x) = f(y|x)\, f(z|x)\, f(x)$$

This assumption is particularly strong and seldom holds in practice. If the CI assumption does not hold then results of SM derived under it will not be valid.

3.1 Parametric macro approaches

Let assume that (X,Y,Z) follow a trivariate normal distribution with parameters:

$$\boldsymbol{\mu} = \begin{pmatrix} \mu_X \\ \mu_Y \\ \mu_Z \end{pmatrix} \qquad \boldsymbol{\Sigma} = \begin{pmatrix} \sigma_X^2 & \sigma_{XY} & \sigma_{XZ} \\ \sigma_{XY} & \sigma_Y^2 & \sigma_{YZ} \\ \sigma_{XZ} & \sigma_{YZ} & \sigma_Z^2 \end{pmatrix}$$

Moreover, let assume that *A* is a sample of n_A i.i.d. observations in which are available just the two variables *X* and *Y*; similarly *B* is a sample of n_B i.i.d. observations containing *X* and *Z*. In this framework, the parameters μ_X and σ_X^2 could be estimated on *A*, on *B* or better on the sample $A \cup B$, obtained by concatenating *A* and *B*. The parameters μ_Y and σ_{XY} can be estimated on *A*, while μ_Z and σ_{XZ} can be estimated on *B*. It remains the covariance σ_{YZ} which cannot be directly estimated because *Y* and *Z* are not jointly observed in the available data sources. One possibility to derive an estimate of such a covariance is provided by assuming the conditional independence of *Y* and *Z* given the *X* variable; in fact, in such a case it comes out that:

$$\sigma_{YZ} = \frac{\sigma_{XY}\sigma_{XZ}}{\sigma_X^2}$$

If the correlation matrix it is considered, the CI assumption imply the following result:

$$\rho_{YZ} = \rho_{XY}\rho_{XZ}$$

and consequently the partial correlation coefficient between *Y* and *Z* conditioning on *X*

$$\rho_{YZ|X} = \frac{\rho_{YZ} - \rho_{XY}\rho_{XZ}}{\sqrt{1-\rho_{XY}^2}\sqrt{1-\rho_{XZ}^2}}$$

becomes equal to 0 ($\rho_{YZ|X} = 0$).

The following expressions show how to estimate the various parameters by considering the sample counterparts and exploiting all the available information as suggested in Kadane (1978) and Moriarity & Scheuren (2001):

$$\hat{\mu}_X = \bar{x}_{A\cup B} = \frac{n_A\bar{x}_A + n_B\bar{x}_B}{n_A + n_B}; \qquad \hat{\sigma}_X^2 = s_{X,A\cup B}^2 = \frac{(n_A-1)s_{X,A}^2 + (n_B-1)s_{X,B}^2}{n_A + n_B - 1}$$

$$\hat{\mu}_Y = \bar{y}_A = \frac{1}{n_A}\sum_{l=1}^{n_A} y_l; \qquad \hat{\sigma}_Y^2 = s_{X,A}^2 = \frac{1}{n_A-1}\sum_{a=1}^{n_A}(y_a - \bar{y}_A)^2$$

$$\hat{\mu}_Z = \bar{z}_B = \frac{1}{n_B}\sum_{b=1}^{n_B} z_b; \qquad \hat{\sigma}_Z^2 = s_{Z,B}^2 = \frac{1}{n_B-1}\sum_{b=1}^{n_B}(z_b - \bar{z}_B)^2$$

$$\hat{\sigma}_{XY} = s_{XY,A} = \frac{1}{n_A-1}\sum_{a=1}^{n_A}(x_a - \bar{x}_A)(y_a - \bar{y}_A)$$

$$\hat{\sigma}_{XZ} = s_{XZ,B} = \frac{1}{n_B-1}\sum_{b=1}^{n_B}(x_b - \bar{x}_B)(z_b - \bar{z}_B)$$

Finally under the CI, it comes out:

$$\hat{\sigma}_{YZ} = \frac{s_{XY,A}\, s_{XZ,B}}{s_{X,A\cup B}^2},$$

To sum up, the estimated variance-covariance matrix would be:

$$\hat{\Sigma} = \begin{pmatrix} s_{X,A\cup B}^2 & s_{XY,A} & s_{XZ,B} \\ s_{XY,A} & s_{Y,A}^2 & \hat{\sigma}_{YZ} \\ s_{XZ,B} & \hat{\sigma}_{YZ} & s_{Z,B}^2 \end{pmatrix}$$

The following example show how to apply these estimation procedures in R with the functions made available by **StatMatch**. In particular, the function `mixed.mtc`, originally developed to perform mixed SM, can be used. In fact when its argument `micro=FALSE` it return just estimates of the parameter of multivariate normal distribution. The parameters are estimated according to the procedure suggested by Moriarity & Scheuren (2001) by setting the argument `method="MS"`.

```
R> # parameters estimated using MS
R> rho.xy <- cor(data.A)
R> rho.xz <- cor(data.B)
R> rho.yz <- rho.xy["x","y"] * rho.xz["x", "z"]  # CI estimate
R> rho.yz
[1] 0.4957038

R> install.packages("StatMatch") # install StatMatch
R> library(StatMatch)  # load StatMatch

# estimation of the parameters of the multivariate normal under CIA
R> mix.MS <- mixed.mtc(data.rec=data.A, data.don=data.B, match.vars="x",
R+                     y.rec="y", z.don="z",
R+                     method="MS", rho.yz=rho.yz, micro=FALSE)
input value for rho.yz is 0.4957038
low(rho.yz)= 0.1382727
up(rho.yz)= 0.8531349
The input value for rho.yz is admissible

R> names(mix.MS)
[1] "rho.yz"  "mu"      "vc"      "cor"     "phi"     "res.var" "call"

R> mix.MS$rho.yz
    start   low.lim    up.lim      used
0.4957038 0.1382727 0.8531349 0.4957038

R> mix.MS$mu #estimated means
          x           y           z
 0.01830470  0.01748767 -0.02698794

R> mix.MS$vc #estimated var-cov matrix
          x         y         z
x 1.2995727 0.4902874 1.3051727
y 0.4902874 0.8877026 0.5420962
z 1.3051727 0.5420962 1.3472267

R> mix.MS$cor #estimated cor. matrix
          x         y         z
x 1.0000000 0.4564746 0.9863870
y 0.4564746 1.0000000 0.4957038
z 0.9863870 0.4957038 1.0000000
```

This way of working is simple and fast but may provide incoherent estimates. In fact, the estimates from different independent samples may lead to a variance-covariance matrix $\hat{\boldsymbol{\Sigma}}$ not positive semidefinite (i.e. the determinant is negative) which is a fundamental property of

the variance-covariance matrix. To avoid such a problem it is possible to resort to Maximum Likelihood (ML) estimation for partially observed data (cf. Section 2.1.1 in D'Orazio *et al.* 2006b).

The ML estimates of the mean and the variance the *X* variable are derived from the concatenated file:

$$\hat{\mu}_X = \overline{x}_{A\cup B}, \qquad \hat{\sigma}^2_X = s^2_{X,A\cup B} = \frac{n_A s^2_{X,A} + n_B s^2_{X,B}}{n_A + n_B},$$

Vice versa, the parameters involving *Y* are estimated by considering the regression equation $Y = \alpha_Y + \beta_{YX} X + \varepsilon_{Y|X}$. Slope and intercept are estimated by (note that variances and covariances here are estimated using ML):

$$\hat{\beta}_{YX} = s_{XY;A} \big/ s^2_{X;A}, \qquad \hat{\alpha}_Y = \overline{y}_A - \hat{\beta}_{YX}\overline{x}_A$$

and, consequently

$$\hat{\mu}_Y = \hat{\alpha}_Y + \hat{\beta}_{YX}\hat{\mu}_X .$$

By following the same reasoning it is possible to derive an estimate of the variance of *Y*:

$$\hat{\sigma}^2_Y = s^2_{Y,A} + \hat{\beta}_{YX}\left(s^2_{X,A\cup B} - s^2_{X,A}\right),$$

while the covariance is obtained in the following manner

$$\hat{\sigma}_{XY} = \hat{\beta}_{YX} s^2_{X,A\cup B} .$$

The same happens as far *Z* is concerned:

$$\hat{\mu}_Z = \hat{\alpha}_Z + \hat{\beta}_{ZX}\hat{\mu}_X, \quad \hat{\sigma}^2_Z = s^2_{Z,B} + \hat{\beta}_{ZX}\left(s^2_{X,A\cup B} - s^2_{X,B}\right), \quad \hat{\sigma}_{XZ} = \hat{\beta}_{ZX} s^2_{X,A\cup B}$$

with $\hat{\alpha}_Z = \overline{z}_B - \hat{\beta}_{ZX}\overline{x}_B$ and $\hat{\beta}_{ZX} = s_{ZY;B} \big/ s^2_{X;B}$.

In R the these ML estimates of the parameters are returned by `mixed.mtc` by setting `method="ML"` and `micro=FALSE`, as shown in the following example.

```
R> # parameters estimated using ML
R> mix.ML <- mixed.mtc(data.rec=data.A, data.don=data.B, match.vars="x",
R+                     y.rec="y", z.don="z",
R+                     method="ML", rho.yz=rho.yz, micro=FALSE)

R> mix.ML$mu #estimated means
          x           y           z
 0.01830470 -0.05120828  0.03955610

R> mix.ML$vc #estimated var-cov matrix
```

```
          x         y         z
x 1.2895760 0.7019641 1.0879601
y 0.7019641 0.9905097 0.7828291
z 1.0879601 0.7828291 1.1608974

R> mix.ML$cor #estimated cor. matrix
          x         y         z
x 1.0000000 0.6211008 0.8891859
y 0.6211008 1.0000000 0.7300300
z 0.8891859 0.7300300 1.0000000
```

Let consider the case of categorical variables; the parameters of interest are the probabilities:

$$\theta_{ijk} = \Pr\left(X = i, Y = j, Z = k\right), \quad i = 1,\ldots,I, \qquad j = 1,\ldots,J, \quad k = 1,\ldots,K$$

with $0 \le \theta_{ijk} \le 1$ and $\sum_{i,j,k} \theta_{ijk} = 1$. Under the CI assumption, it is possible to write the joint probabilities as the result of the following factorization:

$$\Pr\left(X = i, Y = j, Z = k\right) = \Pr\left(Y = j \middle| X = i\right) \Pr\left(Z = k \middle| X = i\right) \Pr\left(X = i\right)$$

and consequently:

$$\theta_{ijk} = \theta_{j|i}\, \theta_{k|i}\, \theta_{i++} = \frac{\theta_{ij+}}{\theta_{i++}} \frac{\theta_{i+k}}{\theta_{ij+}} \theta_{i++} = \frac{\theta_{ij+}\, \theta_{i+k}}{\theta_{i++}}, \qquad i = 1,\ldots,I; \quad j = 1,\ldots,J, \quad k = 1,\ldots,K$$

The probabilities of the marginal table $Y \times Z$ are obtained by summing over *i*:

$$\sum_i \theta_{ijk} = \sum_{i=1}^{I} \frac{\theta_{ij+}\, \theta_{i+k}}{\theta_{i++}}, \qquad j = 1,\ldots,J, \quad k = 1,\ldots,K$$

If $n_{A,ij+}$ are the frequencies in the table $X \times Y$ built using data in *A* and $n_{B,i+k}$ are the counts in the table $X \times Z$ derived from *B*, by applying the ML estimation method it comes out:

$$\hat{\theta}_{ijk} = \frac{\hat{\theta}_{ij+}\, \hat{\theta}_{i+k}}{\hat{\theta}_{i++}}$$

with

$$\hat{\theta}_{i++} = \frac{n_{A,i++} + n_{B,i++}}{n_A + n_B}, \qquad i = 1,\ldots,I$$

$$\hat{\theta}_{j|i} = \frac{n_{A,ij+}}{n_{A,i++}}, \qquad i = 1,\ldots,I, \qquad j = 1,\ldots,J$$

$$\hat{\theta}_{k|i} = \frac{n_{B,i+k}}{n_{B,i++}}, \qquad i = 1,\ldots,I, \qquad k = 1,\ldots,K.$$

The following example shows how to estimate the contingency table of $Y \times Z$ in R with the function `Frechet.bounds.cat` in the package **StatMatch**. In particular this function requires the estimated frequencies (absolute or relative) for the *X* variables (argument `tab.x`), the frequencies in the table $X \times Y$ (argument `tab.xy`) and the frequencies in the table $X \times Z$ (argument `tab.xz`). The data set used are the artificial ones generated in R by means of the code reported in the Appendix A.2.

```
R> intersect(names(rec.A),names(don.B)) # common variables
[1] "hsize"  "hsize6" "db040"  "age"    "c.age"  "rb090"  "pb220a" "rb050"
R>
R> # joint distribution of the X vars: gender vs. age
R> tx.A <- xtabs(~rb090+c.age, data=rec.A)
R> tx.B <- xtabs(~rb090+c.age, data=don.B)
R> tx <- tx.A+tx.B
R> tx
        c.age
rb090    [16,24] (24,49] (49,64] (64,100]
  male       700    2128    1058      727
  female     659    2205    1080      965
R>
R> txy.A <- xtabs(~rb090+c.age+work, data=rec.A)
R> txz.B <- xtabs(~rb090+c.age+c.netI, data=don.B)
R>
R> out <- Frechet.bounds.cat(tab.x=tx, tab.xy=txy.A, tab.xz=txz.B)
R> out$CIA # table of Y (work) vs. Z (classes of netIncome) under CI
             c.netI
work               (-6,0]       (0,5]      (5,10]     (10,15]     (15,20]
  working     0.066361749 0.048019106 0.065171910 0.083980797 0.104412772
  not working 0.062929380 0.044856451 0.074919314 0.089177796 0.081376131
             c.netI
work              (20,25]     (25,30]     (30,40]     (40,50]    (50,200]
  working     0.078680359 0.042240470 0.032957951 0.008151041 0.009910382
  not working 0.054174545 0.023914668 0.018112708 0.005022875 0.005629595
```

3.2 Parametric micro approaches

Different methods are available depending on the nature of the variables (categorical, continuous or mixed) and the "origin" of the synthetic data set, i.e. if obtained by concatenating *A* and *B* or just by considering one of them as the recipient.

When dealing with continuous variables a very common method is the *conditional mean matching* which is based on regression imputation (cf. D'Orazio *et al.*, 2006b, Sec. 2.2.1). The methods is simple and can be used regardless the origin of the synthetic data set is derived. When the synthetic file is obtained by considering *A* as the recipient, then it is filled in with the following predicted values:

$$\hat{z}_a^{(A)} = \hat{\alpha}_Z + \hat{\beta}_{ZX} x_a, \qquad a = 1,2,\ldots,n_A$$

where $\hat{\beta}_{ZX} = s_{XZ;B} / s^2_{X;B}$ and $\hat{\alpha}_Z = \overline{z}_B - \hat{\beta}_{ZX}\overline{x}_A$.

When the synthetic file is obtained by concatenating *A* and *B*, then *A* imputed with $\hat{z}_a^{(A)}$ has be joined with data set *B* is filled in with the following predicted values:

$$\hat{y}_b^{(B)} = \hat{\alpha}_Y + \hat{\beta}_{YX} x_b, \qquad b = 1,2,\ldots,n_B$$

with $\hat{\alpha}_Y = \overline{y}_A - \hat{\beta}_{YX}\overline{x}_B$, and $\hat{\beta}_{YX} = s_{XY;A} / s^2_{X;A}$.

Such regression imputation is relatively simple but unfortunately provides "artificial values" (i.e. values not really observed) lying on the regression line without variation around it. To preserve variability Little and Rubin (2002) suggest to add a random residual to each predicted value (*stochastic regression imputation*). Therefore, *A* is filled in with the values:

$$\tilde{z}_a^{(A)} = \hat{z}_a^{(A)} + e_a = \hat{\alpha}_Z + \hat{\beta}_{ZX} x_a + e_a, \qquad a = 1,2,\ldots,n_A$$

being with e_a is a residual generated randomly from $\mathrm{N}\left(0, \hat{\sigma}_{Z|X}\right)$ with $\hat{\sigma}^2_{Z|X} = s^2_{Z,B} - \hat{\beta}^2_{ZX} s^2_{X,B}$. In the same manner, if necessary, *B* is filled in with the values:

$$\tilde{y}_b^{(B)} = \hat{y}_b^{(B)} + e_b = \hat{\alpha}_Y + \hat{\beta}_{YX} x_b + e_b, \qquad b = 1,2,\ldots,n_B$$

and e_b is a residual generated randomly from $\mathrm{N}\left(0, \hat{\sigma}_{Y|X}\right)$ with $\hat{\sigma}^2_{Y|X} = s^2_{Y,A} - \hat{\beta}^2_{YX} s^2_{X,A}$.

Kadane (1978) showed that the conditional mean matching has the further problem of leading to underestimation of the covariance between *Y* and *Z*. Such underestimation does not happen with stochastic regression imputation where the final synthetic data set can be considered "representative" of the chosen probability density function.

The following R code shows how to perform stochastic regression imputation in R by using "basic" functions (**StatMatch** or other additional packages are not needed).

```
R> A <- as.data.frame(data.A)
R> B <- as.data.frame(data.B)
R>
R> # regression Y vs. X in A
R> reg.yx <- lm(y~x, data=A)
R> coefficients(reg.yx)
(Intercept)           x
  0.3672843   0.5205207
R> out <- summary(reg.yx)
R> out$sigma #residual sd s_Y|X
[1] 0.7703143
R>
R> pred.y.B <- predict(reg.yx, newdata=B) #predicted values
R> imp.y.B <- pred.y.B + rnorm(nrow(B), mean=0, sd=out$sigma)
R>
R> # fill in Y in B
R> B$y <- imp.y.B
R>
R> # regression YZvs. X in B
```

```
R> reg.zx <- lm(z~x, data=B)
R> coefficients(reg.zx)
(Intercept)           x
 0.03515444  0.87254856
R> out <- summary(reg.zx)
R> out$sigma #residual sd s_Z|X
[1] 0.5256025
R>
R> pred.z.A <- predict(reg.zx, newdata=A) #predicted values
R> imp.z.A <- pred.z.A + rnorm(nrow(A), mean=0, sd=out$sigma)
R>
R> # fill in Z in A
R> A$z <- imp.z.A
R>
R> # concatenation A U B
R> AuB <- rbind(A,B)
R> head(AuB)
            x          y          z
1 -1.47011062  0.2231917 -1.6599483
2  0.22957161  2.6339102  0.8565873
3  0.04839709 -0.9714869  0.5625099
4  0.84314453  1.4781465  0.6227054
5  1.71006937  1.9395058  1.9897759
6  0.01294460  0.1524463 -0.2343014
R>
R> cor(AuB) #estimated var-cov
          x         y         z
x 1.0000000 0.5356676 0.8712686
y 0.5356676 1.0000000 0.4933929
z 0.8712686 0.4933929 1.0000000
R> mix.ML$cor
          x         y         z
x 1.0000000 0.6211008 0.8891859
y 0.6211008 1.0000000 0.7300300
z 0.8891859 0.7300300 1.0000000
```

3.3. Comments on the parametric approach

The SM parametric methods require necessarily the specification of a model; then parameters have to be estimated: the choice of the estimation method requires additional evaluations due to the particular SM framework.

Such approach can become too burdensome when dealing with many variables and, as usually happens, of mixed type (categorical and continuous). In some cases some transformations may help: i.e. substitute a categorical variables with dummies and treat them as continuous variables (cannot be applied when there are rare categories) or categorize the continuous variables. Such operation is not straightforward and may introduce undesired noise. In general, the main problem relies in a wrong specification of the model: it is likely to provide unreliable results.

3.4 Nonparametric micro approaches

Nonparametric approaches to SM do not explicitly refer to a model. Most of them consist in filling in the data set chosen as the recipient with the values of the variable which is available only in the other data set, the donor one. In the following it will be assumed that *A* is the recipient while *B* is the donor and $n_A << n_B$; the objective of SM will be that of filling in *A* with values of *Z* (variable available only in *B*).

Some of the proposed SM nonparametric micro methods consist in extending *hot deck imputation* methods developed to impute the missing values in survey data. In particular, commonly used SM methods are (see Section 2.4 in D'Orazio *et al.*, 2006b; Singh *et al.*, 1993):

- *random hot deck*
- *distance hot deck*
- *rank hot deck*

3.4.1 Random hot deck

For each record in *A*, a donor record in *B* is randomly selected and the value of *Z* observed in it is imputed in *A*. Usually, before the random selection, the units in both the dataset are grouped into homogeneous groups (*donation classes*) according to the values of one or more categorical variables, X_G, chosen among the available common variables ($X_G \subseteq X$; for instance gender, region etc.). Then, for a given record in *A* belonging to a given group (say males), it is randomly chosen a donor record in *B* in the same group (males in *B*). In this procedure, a unit in *B* can be selected as a donor more than once.

Such a way of working is equivalent to estimating the conditional distribution of *Z* given X_G (assumed categorical) and drawing an observation from it. In particular, when *Z* is continuous and X_G is categorical, the conditional distribution is estimated via the empirical cumulative distribution:

$$\hat{F}_{Z|X} = \frac{\sum_{b=1}^{n_B} I\left(z_b < z\right) I\left(x_{G,b} = i\right)}{\sum_{b=1}^{n_B} I\left(x_{G,b} = i\right)}.$$

Where $I(\ ) = 1$ if the condition in the parenthesis is satisfied and 0 otherwise. When *Z* is a categorical variable, the conditional distribution is estimated by:

$$\hat{\theta}_{k|i} = \frac{\sum_{b=1}^{n_B} I\left(z_b = k\right) I\left(x_{G,b} = i\right)}{\sum_{b=1}^{n_B} I\left(x_{G,b} = i\right)}$$

The random hot deck matching is performed by the function `RANDwNND.hotdeck` of the package **StatMatch** for the R environment. It carries out the random selection of each donor from a suitable subset of the available donors. This subset can be formed in different

ways. The “basic” random hot deck within imputation classes is performed by simply specifying the donation classes via the argument `don.class` (the classes are formed by crossing the categories of the categorical variables being considered). The following example provides an example of application of random hot deck within classes with the artificial data introduced in the Appendix A.2. The imputation classes are formed by crossing region (variable named `"db040"`) and gender (`"rb090"`).

```
R> group.v <- c("db040","rb090") # var. to form imputation classes
R> rnd.1 <- RANDwNND.hotdeck(data.rec=rec.A, data.don=don.B,
R+                           match.vars=NULL, don.class=group.v)
```

It is worth noting that the function `RANDwNND.hotdeck` does not create the synthetic data set; the identifiers of each recipient record and the corresponding donor are saved in the component `mtc.ids` of the list returned in output. The number of donors available in each donation class are saved in the component `noad`.

```
> head(rnd.1$mtc.ids)# first 6 rows
     rec.id don.id
[1,] "401"  "133"
[2,] "71"   "42"
[3,] "92"   "167"
[4,] "225"  "293"
[5,] "364"  "66"
[6,] "370"  "219"
```

In order to derive the synthetic data set it is necessary to pass the output of `RANDwNND.hotdeck` (component `mtc.ids`) to the function `create.fused`, as shown in the following example.

```
R> fA.rnd <- create.fused(data.rec=rec.A, data.don=don.B,
R+                        mtc.ids=rnd.1$mtc.ids,
R+                        z.vars=c("netIncome", "c.netI"))

R> head(fA.rnd1) # first six obs in the synythetic
    hsize hsize6      db040 age    c.age rb090 pb220a    rb050 pl030
401     5      5 Burgenland  45  (24,49]  male     AT 4.545916     1
71      2      2 Burgenland  65 (64,100]  male     AT 6.151409     5
92      2      2 Burgenland  81 (64,100]  male     AT 6.151409     5
225     3      3 Burgenland  51  (49,64]  male     AT 5.860364     1
364     4      4 Burgenland  18  [16,24]  male     AT 6.316554     1
370     5      5 Burgenland  50  (49,64]  male     AT 4.545916     1
          work      wwA netIncome   c.netI
401     working 10.85782  21407.86  (20,25]
71  not working 14.69250  19092.36  (15,20]
92  not working 14.69250  12442.49  (10,15]
225     working 13.99734  22524.16  (20,25]
364     working 15.08694  53804.59 (50,200]
370     working 10.85782  20890.53  (20,25]
```

The function `RANDwNND.hotdeck` implements various alternative methods to restrict the subset of the potential donors where to select randomly one of them. These methods permits to handle continuous X_M variables ($X_M \subseteq X$) and are based essentially on distance computed on them. For instance, it is possible to consider as potential donors just the units which are below a certain distance to the recipient units according to a distance computed on X_M (cf. Andridge & Little, 2010):

$$d_{ab}\left(x_{M,a}, x_{M,b}\right) \leq \delta; \qquad \delta > 0, \;\; b = 1, 2, \ldots, n_B$$

The variables X_M are passed to the function via the argument `match.vars`. In practice, when `cut.don="k.dist"` only the donors whose distance from the recipient is less or equal to threshold `k` are considered. By setting `cut.don="exact"` the `k` ($0 < k \leq n_D$) closest donors are retained (n_D is the number of available donors for a given recipient). With `cut.don="span"` a proportion `k` ($0 < k \leq 1$) of the closest available donors it is considered while; setting `cut.don="rot"` and `k=NULL` the subset reduces to the $\left[\sqrt{n_D}\right]$ closest donors; finally, when `cut.don="min"` only the donors at the minimum distance from the recipient are retained.

In the following example code, each donor is randomly selected in the subset of the closest $k = 20$ donors in terms of age (`"age"`), sharing the same gender (`"rb090"`) and living in the same region (`"db040"`).

```
R> # random choiches of a donor among the closest k=20 wrt age
R> group.v <- c("db040","rb090")
R> X.mtc <- "age"
R> rnd.2 <- RANDwNND.hotdeck(data.rec=rec.A, data.don=don.B,
R+                           match.vars=X.mtc, don.class=group.v,
R+                           dist.fun="Manhattan",
R+                            cut.don="exact", k=20)
R> fA.knnd <- create.fused(data.rec=rec.A, data.don=don.B,
R+                         mtc.ids=rnd.2$mtc.ids,
R+                         z.vars=c("netIncome", "c.netI"))
```

When distances are computed on some matching variables, then the output of `RANDwNND.hotdeck` provides some information concerning the distances of the possible available donors for each recipient observation.

```
R> head(rnd.2$sum.dist)
     min max       sd cut dist.rd
[1,]   0  47 11.02087   5       3
[2,]   0  49 14.54555   4       1
[3,]   0  65 19.01027   9       4
[4,]   1  41 10.09283   6       1
[5,]   1  74 19.53088  11       6
[6,]   0  42 10.16749   5       4
```

In particular, `"min"`, `"max"` and `"sd"` columns report respectively the minimum, the maximum and the standard deviation of the distances (all the available donors are considered), while `"cut"` refers the maximum distance observed for the donors in the given subset; `"dist.rd"` is distance existing between the recipient and the randomly chosen donor.

When selecting a donor among those available it is possible to use a weighted selection by specifying a weighs via `weight.don` argument. This issue will be tackled in Section 6.

3.4.2 Distance hot deck

Distance hot deck methods are very common in imputation of missing data. In SM, for each record in *A*, it is selected a closest donor record in *B* according to a distance computed on a suitable subset of the common variables X_M ($X_M \subseteq X$)

$$d_{ab}\left(\mathbf{x}_{M,a}, \mathbf{x}_{M,b}\right) = \min, \qquad b = 1, 2, \ldots, n_B$$

then the value of *Z* observed on the donor unit it is imputed in *A*. The choice of the matching variables X_M, that have to be used in computing distances, is a crucial step (for major details see Section 7.2). Usually, they are the subset of the *X* variables that at the same time are connected with *Z* and with *Y*. Choosing too many matching variables may affect negatively the matching results: the marginal distribution of *Z* imputed in *A* may not reflect the one observed in *B*. Many distances can be used to compute proximity between units (see Appendix C in D'Orazio *et al.*, 2006b). Before searching for donors it may be convenient and more efficient to divide units in *A* and *B* into donation classes according to the values of some categorical variables X_G ($X_G \subseteq X$). For instance, for male recipient in *A* it will be searched the closest male donor in *B*.

The nearest neighbour distance hot deck techniques are implemented in the function `NND.hotdeck` in the package **StatMatch**. This function searches in `data.don` the nearest neighbour of each unit in `data.rec` according to a distance computed on the matching variables X_M specified with the argument `match.vars`. By default the Manhattan (city block) distance is considered (`dist.fun="Manhattan"`). In order to reduce the computational effort it is preferable to define some donation classes (argument `don.class`) because in this case the distances are computed only between units belonging to the same class.

In the following, a simple example of usage of `NND.hotdeck` is reported; donation classes are formed using gender and region (`"rb090"`, `"db040"`), while distances are computed on age (`"age"`).

```
R> group.v <- c("rb090","db040")
R> X.mtc <- "age"
R> out.nnd <- NND.hotdeck(data.rec=rec.A, data.don=don.B,
R+                        match.vars=X.mtc, don.class=group.v)
```

The function `NND.hotdeck` does not create the synthetic data set; for each unit in *A* the corresponding closest donor in *B* is identified according to the imputation classes (when defined) and the chosen distance function; the recipient-donor units' identifiers are saved in the data.frame `mtc.ids` stored in the output list returned by `NND.hotdeck`. The output list provides also the distance between each couple recipient-donor (saved in the `dist.rd` component of the output list) and the number of available donors at the minimum distance for each recipient (component `noad`). Note that when there are more donors at the minimum distance, then one of them is picked up at random.

```
R> summary(out.nnd$dist.rd) # summary distances rec-don
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   0.00    0.00    0.00    0.04    0.00    7.00
R> summary(out.nnd$noad) # summary available donors at min. dist.
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   1.00    4.00    6.00    6.56    9.00   21.00
```

In order to derive the synthetic data set it is necessary to run the function `create.fused`:

```
R> head(out.nnd$mtc.ids)
     rec.id don.id
[1,] "401"  "376"
[2,] "71"   "118"
[3,] "92"   "106"
[4,] "225"  "253"
[5,] "364"  "288"
[6,] "370"  "350"

R> fA.nnd <- create.fused(data.rec=rec.A, data.don=don.B,
R+                        mtc.ids=out.nnd$mtc.ids,
R+                        z.vars=c("netIncome","c.netI"))

R> head(fA.nnd) #first 6 obs.
    hsize hsize6      db040 age    c.age rb090 pb220a    rb050 pl030
401     5      5 Burgenland  45  (24,49]  male     AT 4.545916     1
71      2      2 Burgenland  65 (64,100]  male     AT 6.151409     5
92      2      2 Burgenland  81 (64,100]  male     AT 6.151409     5
225     3      3 Burgenland  51  (49,64]  male     AT 5.860364     1
364     4      4 Burgenland  18  [16,24]  male     AT 6.316554     1
370     5      5 Burgenland  50  (49,64]  male     AT 4.545916     1
          work      wwA netIncome  c.netI
401     working 10.85782  47159.21 (40,50]
71  not working 14.69250  21316.32 (20,25]
92  not working 14.69250  21667.53 (20,25]
225     working 13.99734  34166.20 (30,40]
364     working 15.08694  10228.02 (10,15]
370     working 10.85782  20667.61 (20,25]
```

As far as distances are concerned (argument `dist.fun`), all the distance functions in the package **proxy** (Meyer and Butchta, 2013) are available. Anyway, for some particular distances it was decided to write specific R functions. In particular, when dealing with

continuous matching variables it is possible to use the *maximum distance* (L^{∞} norm) implemented in `maximum.dist`; this function works on the true observed values (continuous variables) or on transformed ranked values (argument `rank=TRUE`) as suggested in Kovar *et al.* (1988); the transformation (ranks divided by the number of units) removes the effect of different scales and the new values are uniformly distributed in the interval [0,1]. The Mahalanobis distance can be computed by using `mahalanobis.dist` which allows an external estimate of the covariance matrix (argument `vc`). When dealing with mixed type matching variables, the Gowers's dissimilarity (Gower, 1981) can be computed (function `gower.dist`): it is an average of the distances computed on the single variables according to different rules, depending on the type of the variable. All the distances are scaled to range from 0 to 1, hence the overall distance cat take a value in [0,1].

In traditional distance hot deck a unit in *B* can be chosen more than once as a donor. In order to avoid this, a *constrained distance hot deck* can be used: a donor can be used just once and the subset of the donors is selected in order to minimize the overall matching distance. In this case the selection of the donors requires the solution of an optimization problem (transportation problem). The constrained matching returns an overall matching distance greater than the one in the unconstrained case, but it tends to better preserve the marginal distribution of the variable imputed in the synthetic data set.

In `NND.hotdeck` the multiple usage of a donor can be avoided by setting the argument `constrained=TRUE`; in such a case, a donor can be used just once and all the donors are selected in order to minimize the overall matching distance. In practice, the donors are identified by solving a traveling salesperson problem; two alternatives are available: the Hungarian algorithm (argument `constr.alg="Hungarian"`) implemented in the function `solve_LSAP` in the R package **clue** (Hornik, 2013) and the algorithm provided by the package **lpSolve** (Berkelaar *et al.*, 2013) (argument `constr.alg="lPsolve"`). Setting `constr.alg="Hungarian"` (default) is more efficient and faster.

```
R> group.v <- c("rb090","db040")
R> X.mtc <- "age"
R> out.nnd.c <- NND.hotdeck(data.rec=rec.A, data.don=don.B,
R+                          match.vars=X.mtc, don.class=group.v,
R+                          dist.fun="Manhattan", constrained=TRUE,
R+                          constr.alg="Hungarian")
Warning: The  Manhattan  distance is being used
All the categorical matching variables in rec and don
 data.frames, if present are recoded into dummies
R> fA.nnd.c <- create.fused(data.rec=rec.A, data.don=don.B,
R+                          mtc.ids=out.nnd.c$mtc.ids,
R+                          z.vars=c("netIncome","c.netI"))
R>
R>
R> #comparing distances
R> sum(out.nnd$dist.rd) # unconstrained
[1] 160
R> sum(out.nnd.c$dist.rd) # constrained
[1] 1189
```

3.3.3 Rank hot deck

The rank hot deck distance method has been introduced by Singh *et al.* (1993). It searches for the donor at a minimum distance from the given recipient record but, in this case, the distance is computed on the percentage points of the empirical cumulative distribution function of the unique (continuous) common variable X_M being considered:

$$d_{ab}\left(\hat{F}(x_{M,a}),\ \hat{F}(x_{M,b})\right) = \left|\hat{F}(x_{M,a}) - \hat{F}(x_{M,b})\right|, \qquad b = 1,2,\ldots,n_B$$

$$\hat{F}(x_{M,a}) = \frac{1}{n_A}\sum_{t=1}^{n_A} I\left(x_{M,t} \le x_{M,a}\right), \qquad a = 1,2,\ldots,n_A$$

$$\hat{F}(x_{M,b}) = \frac{1}{n_B}\sum_{t=1}^{n_B} I\left(x_{M,t} \le x_{M,b}\right), \qquad b = 1,2,\ldots,n_B$$

This transformation provides values uniformly distributed in the interval $[0,1]$; moreover, it can be useful when the values of X_M cannot be directly compared because of measurement errors which however do not affect the "position" of a unit in the whole distribution (D'Orazio *et al.*, 2006b).

In **Statmatch** this method is implemented in the function `rankNND.hotdeck`. In it the names of the matching variable in *A* and *B* to be used to compute empirical cumulative distribution should be passed via the argument `var.rec` (name of the variable in `data.rec`) and `var.don` (name of the variable in `data.don`). The following simple example shows how `rankNND.hotdeck` works.

```
R> rnk.1 <- rankNND.hotdeck(data.rec=rec.A, data.don=don.B,
R+                          var.rec="age", var.don="age")
R> #create the synthetic data set
R> fA.rnk <- create.fused(data.rec=rec.A, data.don=don.B,
R+                       mtc.ids=rnk.1$mtc.ids,
R+                       z.vars=c("netIncome", "c.netI"),
R+                       dup.x=TRUE, match.vars="age")
R> head(fA.rnk)
      hsize hsize6         db040 age    c.age  rb090 pb220a    rb050 pl030
4547      2      2     Carinthia  45  (24,49]   male     AT 6.863162     1
9819      4      4      Salzburg  35  (24,49] female     AT 6.089967     1
4461      2      2     Carinthia  57  (49,64]   male     AT 6.863162     1
10222     2      2         Tyrol  69 (64,100] female     AT 6.857877     5
8228      4      4 Upper Austria  25  (24,49] female     AT 6.945309     4
3361      3      3        Vienna  22  [16,24]   male  Other 8.374000     1
             work      wwA age.don netIncome  c.netI
4547      working 16.39250      45  21788.09 (20,25]
9819      working 14.54575      35  19115.94 (15,20]
4461      working 16.39250      58   9921.61  (5,10]
10222 not working 16.37988      70   9747.15  (5,10]
8228  not working 16.58871      25  29027.18 (25,30]
3361      working 20.00110      22  11591.76 (10,15]
```

The function `rankNND.hotdeck` allows for constrained and unconstrained matching in the same manner as in `NND.hotdeck`. It is also possible to define some donation classes (argument `don.class`), in this case the empirical cumulative distribution is estimated separately class by class.

```
R> # constrained rank hot deck with donation classes
R>
R> rnk.2 <- rankNND.hotdeck(data.rec=rec.A, data.don=don.B, var.rec="age",
R+                          var.don="age", don.class="rb090",
R+                          constrained=TRUE, constr.alg="Hungarian")
R> fA.grnk <- create.fused(data.rec=rec.A, data.don=don.B,
R+                         mtc.ids=rnk.2$mtc.ids,
R+                         z.vars=c("netIncome", "c.netI"),
R+                         dup.x=TRUE, match.vars="age")
R> head(fA.grnk)
     hsize hsize6         db040 age   c.age rb090 pb220a    rb050 pl030
4547     2      2     Carinthia  45 (24,49]  male     AT 6.863162     1
4461     2      2     Carinthia  57 (49,64]  male     AT 6.863162     1
3361     3      3        Vienna  22 [16,24]  male  Other 8.374000     1
827      2      2 Lower Austria  57 (49,64]  male     AT 6.913897     1
8061     3      3 Upper Austria  31 (24,49]  male     AT 7.509383     1
1925     4      4 Lower Austria  49 (24,49]  male     AT 7.757150     1
        work      wwA age.don netIncome   c.netI
4547 working 16.39250      46  23149.70  (20,25]
4461 working 16.39250      59  45463.71  (40,50]
3361 working 20.00110      22  30458.38  (30,40]
827  working 16.51368      59  53567.80 (50,200]
8061 working 17.93599      31  15863.65  (15,20]
1925 working 18.52777      51  56824.81 (50,200]
```

In estimating the empirical cumulative distribution it is possible to consider the units' weights (arguments `weight.rec` and `weight.don`). This topic will be tackled in Section 7.

*3.3.4. Using functions in **StatMatch** to impute missing values in a survey*

All the functions in **StatMatch** that implement the hot deck imputation techniques can be used to impute missing values in a single data set. In this case it is necessary to:

i) separate the observations in two data sets: the file *A* plays the role of recipient and will contain the units with missing values on the target variable, while the file *B* is the donor and will contain all the available donors (units with non-missing values for the target variable).

ii) Fill in the missing values in the recipient, e.g. by using a nonparametric imputation

iii) Join recipient and donor file

3.5 Mixed methods

A SM mixed method consists of two steps: (1) a model is fitted and all its parameters are estimated; and (2) a nonparametric approach is used to create the synthetic data set. The model is more parsimonious while the nonparametric approach offers "protection" against model misspecification. The proposed mixed approaches for SM are based essentially on *predictive mean matching* imputation methods (see D'Orazio *et al.* 2006b, Sections 2.5 and 3.6). In the case of continuous *X*, *Y* and *Z*, a general procedure consists of the following steps:

Step 1) A regression model is assumed as far as *Z* is concerned e.g.

$$Z = g(X;\theta)$$

Its parameters (θ) are estimated. The estimated model is used to derive "artificial" values, $\tilde{z}_a$, of *Z* in *A* (predicted values, or predicted plus a random error term).

Step 2) For each record in *A* it is selected the closest record in *B* according to a distance $d_{ab}(\tilde{z}_a, z_b)$ computed considering artificial and truly observed values of *Z*.
The closest record in *B* donates to *A* the value of *Z* observed on it.

Similar procedures are proposed by Rubin (1986), Singh *et al.* (1993), Moriarity & Scheuren (2001 and 2003).

A further advantage of such a procedure is that it avoid to computing the distances on several common variables, whereas variables with low predictive power on the target variable may influence negatively the distances.

D'Orazio *et al.* (2006b) suggested the following mixed procedure (MM5) when dealing with continuous variables:

Step 1) Two regression models are considered:

$$Y = \alpha_Y + \beta_{YX} X + \varepsilon_Y$$
$$Z = \alpha_Z + \beta_{ZX} X + \varepsilon_Z$$

Their parameters are estimated and then *A* is filled in with the values:

$$\tilde{z}_a^{(A)} = \hat{z}_a^{(A)} + e_a = \hat{\alpha}_Z + \hat{\beta}_{ZX} x_a + e_a, \qquad a = 1, 2, \ldots, n_A$$

being e_a a residual generated randomly from $\mathrm{N}\left(0, \hat{\sigma}_{Z|X}\right)$; in the same manner *B* is filled in with the values:

$$\tilde{y}_b^{(B)} = \hat{y}_b^{(B)} + e_b = \hat{\alpha}_Y + \hat{\beta}_{YX} x_b + e_b, \qquad b = 1, 2, \ldots, n_B$$

being e_b a residual generated randomly from $\mathrm{N}\left(0, \hat{\sigma}_{Y|X}\right)$.

Step 2) For each record in *A* it is selected the closest record in *B* according to the distance

$$d_{ab}\left(\left(y_a, \tilde{z}_a\right), \left(\tilde{y}_b, z_b\right)\right).$$

The Mahalanobis distance is applied and the matching is constrained.

The parameter of the regression model can be estimated by using ML (as shown in Section 3.1) or the procedure proposed by Moriarity and Scheuren (2001, 2003). D'Orazio *et al.* (2005) compared the mixed procedure based on the two alternative methods for estimating the parameters in an extensive simulation study; in general ML tends to perform better, moreover it permits to avoid some incoherencies in the estimation of the parameters that can happen with the Moriarity and Scheuren approach. Both the mixed procedures are made available by the function `mixed.mtc`. By default, ML (argument `method="ML"`) is applied in the step (1); the method proposed by Moriarity & Scheuren is used when `method="MS"`.

In the following example the `iris` data set is used just to show how `mixed.mtc` works.

```
R> mix.MLm <- mixed.mtc(data.rec=data.A, data.don=data.B, match.vars="x",
R+                      y.rec="y", z.don="z", method="ML", rho.yz=0,
R+                       micro=TRUE, constr.alg="Hungarian")
```

It is worth noting that `mixed.mtc` provides directly the synthetic data set as the component `filled.rec` of the output list returned by calling it with the argument `micro=TRUE`..

```
R> head(mix.MLm$filled.rec) #synthetic data set
             x          y          z
[1,]  0.9059591 -0.2773861  0.6095526
[2,]  2.0538185  0.6624680  1.3799476
[3,]  1.3294165  0.6900273  1.1860791
[4,]  0.6913546 -0.6068382  1.1600654
[5,]  0.5166972  0.2516051  1.0994776
[6,] -1.0335245 -1.6235987 -0.9229601

R> mix.MLm$mu # estimated means via ML
           x           y           z
-0.046662335 -0.004448075  0.007677143

R> colMeans(mix.MLm$filled.rec) # means estimated on the synt. data set
         x          y          z
-0.2838943 -0.1309391 -0.1432121
```

```
R> mix.MLm$cor # estimated correlations via ML
          x         y         z
x 1.0000000 0.5727024 0.8767873
y 0.5727024 1.0000000 0.5021382
z 0.8767873 0.5021382 1.0000000

R> cor(mix.MLm$filled.rec) # correlations estimated on the synt. data set
          x         y         z
x 1.0000000 0.5620894 0.8650576
y 0.5620894 1.0000000 0.5011865
z 0.8650576 0.5011865 1.0000000
```

4. Statistical matching with auxiliary information

In the traditional SM framework is not possible to test whether the conditional independence assumption holds or not; when it is not valid, the SM results obtained by assuming it are biased. In order to obtain reliable results it would be advisable to use some *auxiliary information* in the SM. The term auxiliary information here is used in a wider meaning referring to one or more of the following elements:

a) a third data source *C* where (*X*, *Y*, *Z*) or just (*Y*, *Z*) are jointly observed (e.g. small survey, past survey or census data, administrative register, etc.).

b) Estimates related to the parameters of (Y,Z) or $(Y,Z)|X$ (e.g. estimate of the covariance σ_{YZ}; of the correlation coefficient ρ_{YZ} or of the partial correlation coefficient $\rho_{YZ|X}$; a contingency table of $Y \times Z$; etc.).

c) A priori knowledge of the investigated phenomenon which allow to identify some logical constraints on the parameters' values. For instance, when dealing with categorical variables there may be some *structural zeros*, i.e. events that cannot happen because they are impossible (cf. Agresti, 2002, p. 392); for instance:

Pr(Age=6 AND Education Level="Univ. Degree") = 0

4.1 Parametric macro objective: use of an additional data source

Let assume that the auxiliary information consists in an additional data source *C* containing n_C i.i.d. observations. In this case the parameter can be estimated on the file obtained by concatenating all the data sources ($A \cup B \cup C$) and then by exploiting all the available information. In particular, when all the variables (*X*, *Y*, *Z*) are available in *C* then some parameters can be estimated in closed form; while iterative methods (e.g. EM) are necessary to estimate parameters related to $(Y,Z)|X$ (information available just in *C*). On the contrary, when just *Y* and *Z* are available in *C* (*X* is missing) then iterative methods (e.g.

EM) are necessary to estimate all the parameters. For major details see D'Orazio *et al.* (2006b, pp. 68-71).

4.2 Parametric macro: use of an external estimate

In some cases external estimates are available from previous surveys or from external agencies. The estimates of interest here are those concerning the parameters that cannot be estimated from the available data sources.

Unfortunately, it is not simple to handle an estimation process by including an estimate of a parameter coming from external sources, because it may be not compatible with the available data. For instance, in the case of the trivariate normal distribution, given the available data sources *A* and *B*, an external estimate ρ^*_{YZ} of the correlation coefficient between *Y* and *Z* is compatible with the available data if:

$$\hat{\rho}_{XY}\hat{\rho}_{XZ} - \left[\left(1-\hat{\rho}^2_{XY}\right)\left(1-\hat{\rho}^2_{XZ}\right)\right]^{1/2} \leq \rho^*_{YZ} \leq \hat{\rho}_{XY}\hat{\rho}_{XZ} + \left[\left(1-\hat{\rho}^2_{XY}\right)\left(1-\hat{\rho}^2_{XZ}\right)\right]^{1/2}$$

being $\hat{\rho}_{XY}$ and $\hat{\rho}_{XZ}$ the correlation coefficients estimated from the available data. This result derives by the fact that the correlation matrix has to be positive semidefinite.

To handle non compatible estimates of ρ^*_{YZ} Moriarity & Scheuren (2001, 2003) suggest substituting it with a compatible one. The choice of the compatible value to use into the estimation is left to the analyst.

In **StatMatch**, the estimation of the parameters of the normal distribution by including an external estimates of ρ^*_{YZ}, as suggested by Moriarity & Scheuren (2001, 2003), is implemented by the function `mixed.mtc` (`method="MS", micro=FALSE`). In particular, the estimate should be passed via the argument `rho.yz`. In the following example the data of Appendix A.1 are considered; it is assumed of having an external estimate of the correlation to coefficient between *Y* and *Z* equal to 0.95 ($\rho^*_{YZ} = 0.95$).

```
R> mix.MS <- mixed.mtc(data.rec=data.A, data.don=data.B, match.vars="x",
R+                     y.rec="y", z.don="z",
R+                     method="MS", rho.yz=0.95, micro=FALSE)
input value for rho.yz is 0.95
low(rho.yz)= 0.05119875
up(rho.yz)= 0.9409318
Warning: value for rho.yz is not admissible: a new value is chosen for it
The new value for rho.yz is  0.9309318
R> mix.MS$cor
          x         y         z
x 1.0000000 0.5873950 0.8448829
y 0.5873950 1.0000000 0.9309318
z 0.8448829 0.9309318 1.0000000
```

In the example the external estimate $\rho^*_{YZ} = 0.95$ is not compatible with the available data and it is substituted with a close admissible value, i.e. $\rho^*_{YZ} = 0.9309$ (a value just below the corresponding bound $\rho^{(up)}_{YZ} = 0.9409$).

The situation changes if instead of the correlation coefficient ρ^*_{YZ} it is available the partial correlation coefficient $\rho^*_{YZ|X}$, in fact, in such a case, by using the ML estimation it comes out (cf. D'Orazio *et al.*, 2005):

$$\tilde{\sigma}_{YZ|X} = \rho^*_{YZ|X} \sqrt{\hat{\sigma}^2_{Y|X}\ \hat{\sigma}^2_{Z|X}}$$

It is worth noting that all the values that $\rho^*_{YZ|X}$ can assume ($0 \le \rho^*_{YZ|X} \le 1$) are always compatible with the available data. Unfortunately, in applications, it is difficult that an external estimates of the $\rho^*_{YZ|X}$ it is available. Usage of external estimates of the partial correlation coefficient are shown in Rubin (1986) and Rassler (2002 and 2003). D'Orazio *et al.* (2005) compared how the various methods performs in the case of the multivariate normal distribution. Suggestion from Moriarity & Scheuren are taken into account too. The study shows that methods based on ML estimation based on the usage of the partial correlation coefficient tends to perform better.

```
R> mix.ML <- mixed.mtc(data.rec=data.A, data.don=data.B, match.vars="x",
R+                     y.rec="y", z.don="z",
R+                     method="ML", rho.yz=0.85, micro=FALSE)
R> mix.ML$cor
          x         y         z
x 1.0000000 0.6036985 0.8294118
y 0.6036985 1.0000000 0.8792648
z 0.8294118 0.8792648 1.0000000
```

The problem of compatible external estimates holds in the case of categorical variables too. In particular, if external estimates:

$$\theta^*_{+jk} = \Pr^*\left(Y = j, Z = k\right), \qquad j = 1,\ldots,J,\ k = 1,\ldots,K$$

of the cell probabilities in the contingency table $Y \times Z$ are available, then these estimates are compatible with the available data if the marginal distributions of *Y* and *Z* are equal to the corresponding marginal distribution that can be obtained by the available data sources, i.e.

Marginal distr. of *Y*: $\theta^*_{+j+} = \sum_{k=1}^{K} \theta^*_{+jk} = \sum_{i=1}^{I} \hat{\theta}_{i++} \hat{\theta}_{j|i} \qquad j = 1, 2, \ldots, J$

Marginal distr. of *Z*: $\theta^*_{++k} = \sum_{j=1}^{J} \theta^*_{+jk} = \sum_{i=1}^{I} \hat{\theta}_{i++} \hat{\theta}_{k|i} \qquad k = 1, 2, \ldots, K$

with

$$\hat{\theta}_{i++} = \frac{n_{A,i++} + n_{B,i++}}{n_A + n_B}, \qquad i = 1,\ldots,I$$

$$\hat{\theta}_{j|i} = \frac{n_{A,ij+}}{n_{A,i++}}; \qquad \hat{\theta}_{k|i} = \frac{n_{B,i+k}}{n_{B,i++}} \qquad i = 1,\ldots,I\,, \quad j = 1,\ldots,J\,,\ k = 1,\ldots,K$$

Unfortunately when θ^*_{+jk} are compatible this additional information do not solve completely the estimation problem, because there are no enough information to derive a unique estimate of the θ_{ijk} unless it is considered a loglinear model in which the three-way interaction parameters are set to 0.

If the external information consists in estimates conditional parameters, i.e. probabilities of the conditional distribution of $(Y \times Z)|X$

$$\theta^*_{jk|i} = \Pr^*\left(Y = j, Z = j \,|\, X = i\right), \qquad i = 1,\ldots,I\,,\ j = 1,\ldots,J\,,\ k = 1,\ldots,K$$

and they are reliable, it is possible to plugging in them into the estimation process using data in *A* and *B* just to estimate the marginal distribution of *X*:

$$\theta_{ijk} = \theta^*_{jk|i}\hat{\theta}_{i++} = \theta^*_{jk|i}\,\frac{n_{A,i++} + n_{B,i++}}{n_A + n_B}, \qquad i = 1,\ldots,I\,,\ j = 1,\ldots,J\,,\ k = 1,\ldots,K$$

Unfortunately, in practice it is quite rare that $\theta_{jk|i}$ are known in advance.

4.3 Use of external information in the parametric micro approach

In this case a model is assumed for joint distribution of *X*, *Y* and *Z* and then the parameters are estimated exploiting all the available information (starting data sources and auxiliary information); when the parameters can be estimated uniquely then the estimates are used generate a synthetic data source using *conditional mean matching* or draws from the predicted distribution.

The additional information is used just at the estimation stage but it is no used in the generation of the synthetic data set.

4.4 Use of external information in the nonparametric micro approach

When the objective is the creation of a synthetic data set and the auxiliary information is represented by an additional data source *C*, it is common to apply nonparametric methods. Various alternative procedures can be found in Singh *et al.* (1993).

In case of the distance hot deck different procedures are suggested depending on the "quality" of the auxiliary information. In particular, if *C* is a sample large enough and information provided by it can be considered reliable, then the procedure to apply consists in imputing *Z* in *A* using *C* as donor. Distance between record is computed in the following manner:

- $d_{ac}\left(\left(x_a, y_a\right),\left(x_c, y_c\right)\right)$, if *C* contains *X*, *Y* and *Z*;
- $d_{ac}\left(y_a, y_c\right)$, if *C* contains just *Y* and *Z*

When *C* is a small sample and information in it is not fully reliable, then a two-step procedure should be applied:

Step 1) impute *Z* in *A* using *C* as donor with distance:

- $d_{ac}\left(\left(x_a, y_a\right),\left(x_c, y_c\right)\right)$, if *C* contains *X*, *Y* and *Z*.
- $d_{ac}\left(y_a, y_c\right)$, if *C* contains just *Y* and *Z*

Step 2) impute *Z* in *A* using *B* as a donor with distance computed using:

$$d_{ab}\left(\left(x_a, \tilde{z}_a\right),\left(x_b, z_b\right)\right)$$

being $\tilde{z}_a$ imputed in *A* after the step (1).

When the information in *C* is not fully reliable the two steps "robustify" the procedure; in practice, information in *C* provide an idea of the relationship existing among the variables but the values of the variables are assumed to be affected by errors (measurement errors or outdated values) and therefore they are not directly used for imputation in *A*.

4.5 Use of an external estimate in the mixed micro approach

Let consider the case of the trivariate normal distribution. The procedure MM6 suggested in D'Orazio *et al* (2005) permits to plug in and external estimate of $\rho^{*}_{YZ|X}$ in the step (1) of the mixed procedure when estimating the regression parameters using ML. In particular, the external estimate of the partial correlation coefficient is used to derive the estimate of the partial covariance:

$$\hat{\sigma}_{YZ|X} = \rho^{*}_{YZ|X} \sqrt{\hat{\sigma}^2_{Y|X}\ \hat{\sigma}^2_{Z|X}}$$

and then the covariance between *Y* and *Z*:

$$\hat{\sigma}_{YZ} = \hat{\sigma}_{YZ|X} + \left(\hat{\sigma}_{XY}\hat{\sigma}_{XZ}\right)/\hat{\sigma}^2_X$$

The step (2) of the mixed procedure remains unchanged and consist in

2.a) imputing the following intermediate values:

in file *A*: $\tilde{z}_a = \hat{\mu}_Z + \frac{\hat{\sigma}_{ZX|Y}}{\hat{\sigma}^2_{X|Y}}\left(x_a - \hat{\mu}_X\right) + \frac{\hat{\sigma}_{ZY|X}}{\hat{\sigma}^2_{Y|X}}\left(y_a - \hat{\mu}_Y\right) + e_a, \qquad a = 1,\ldots,n_A$

in file *B*: $\tilde{y}_b = \hat{\mu}_Y + \frac{\hat{\sigma}_{YX|Z}}{\hat{\sigma}^2_{X|Z}}\left(x_b - \hat{\mu}_X\right) + \frac{\hat{\sigma}_{YZ|X}}{\hat{\sigma}^2_{Z|X}}\left(z_b - \hat{\mu}_Z\right) + e_b, \qquad b = 1,\ldots,n_B$

e_a is a random residual drawn from $N\left(0, \hat{\sigma}^2_{Z|XY}\right)$;

e_b is a random residual drawn from $N\left(0, \hat{\sigma}^2_{Y|XZ}\right)$.

2.b) final nonparametric imputation: for each record *a* in file *A*, it is imputed the value z_b observed on a the unit in *B* which is closest according to a Mahalanobis distance computed using observed and intermediate imputed values for *Y* and *Z*:

$d\left[\left(y_a, \tilde{z}_a\right), \left(\tilde{y}_b, z_b\right)\right]$

A constrained matching is applied.

This procedure is implemented in the function `mixed.mtc` by setting `method="ML"` and `micro=TRUE` and passing the guess $\rho^*_{YZ|X}$ through the argument `rho.yz`. In the following example data from Appendix A.1 are considered and it is considered $\rho^*_{YZ|X} = 0.85$

```
R> # mixed procedure
R> # parameters estimated using ML and rho_YZ|X=0.85
R> # and micro=T
R>
R> mix.ML.m <- mixed.mtc(data.rec=data.A, data.don=data.B, match.vars="x",
R+                       y.rec="y", z.don="z",
R+                       method="ML", rho.yz=0.85, micro=TRUE,
R+                       constr.alg="Hungarian")
R> head(mix.ML.m$filled.rec)
             x          y          z
[1,]  0.5988359  1.8481923  1.3546985
[2,]  0.2808349  0.4076143 -0.3159812
[3,]  1.1360676  2.5094471  1.4200685
[4,] -1.7441733 -1.4601703 -1.4324159
[5,]  0.3367595  1.7162756  1.0637088
[6,] -1.5830066 -2.5580300 -2.1198548

R> mix.ML.m$cor
          x         y         z
x 1.0000000 0.6036985 0.8294118
y 0.6036985 1.0000000 0.8792648
z 0.8294118 0.8792648 1.0000000
R> cor(mix.ML.m$filled.rec) # correlation estimated on the synthetic
          x         y         z
x 1.0000000 0.5958282 0.8183102
y 0.5958282 1.0000000 0.8138538
z 0.8183102 0.8138538 1.0000000
```

Moriarity and Scheuren (2001, 2003) proposed a similar procedure that permits to incorporate additional information represented by the correlation coefficient ρ_{YZ}^{*} and where the parameters of the normal distribution are estimated by their sample counterpart. The procedure is implemented in the function `mixed.mtc` (`method="MS"` and `micro=TRUE`).

It is worth noting that `mixed.mtc` permits categorical variables to be included in the set of the matching variables X_M (argument `match.vars`). From the practical viewpoint the categorical variables are substituted by the corresponding dummy variables and the multivariate normal distribution is still assumed. Obviously, such an assumption is rather questionable and care should be used when evaluating the final results.

Singh *et al.* (1993) introduced various mixed methods to deal with the case of categorical variables (cf. D'Orazio *et al.* 2006b, Section 3.6.3 and 3.7).

5. Exploring uncertainty in statistical matching framework

When the objective of SM consists in estimating a parameter (macro approach) it is possible to reason in terms of *uncertainty* on the model chosen for (X_M, Y, Z). Uncertainty is due to the lack of knowledge typical of the basic SM framework (Y and Z are not jointly observed). This approach does not end with a unique estimate of the unknown parameter characterizing the joint p.d.f. for (X_M, Y, Z); on the contrary it identifies an interval of plausible values for it.

When dealing with categorical variables, the intervals of plausible values for the probabilities in the table $Y \times Z$ are derived by means of the Fréchet bounds:

$$\max\{0; P_{Y=j} + P_{Z=k} - 1\} \leq P_{Y=j,Z=k} \leq \min\{P_{Y=j}; P_{Z=k}\} \qquad j = 1,\ldots,J;\ k = 1,\ldots,K$$

being J and K the categories of Y and Z respectively.

Let consider the matching variables X_M, for sake of simplicity let assume that X_D is the variable obtained by the crossproduct of the chosen X_M variables; by conditioning on X_D, it is possible to derive the following result (D'Orazio *et al.*, 2006a):

$$P_{j,k}^{(low)} \leq P_{Y=j,Z=k} \leq P_{j,k}^{(up)}$$

with

$$P_{j,k}^{(low)} = \sum_i P_{X_D=i} \max\left\{0; P_{Y=j|X_D=i} + P_{Z=k|X_D=i} - 1\right\}, \qquad j = 1,\ldots,J;\ k = 1,\ldots,K$$

$$P_{j,k}^{(up)} = \sum_i P_{X_D=i} \min\left\{P_{Y=j|X_D=i}; P_{Z=k|X_D=i}\right\} \qquad j = 1,\ldots,J;\ k = 1,\ldots,K.$$

It is interesting to observe that under the CI assumption the probability $P_{Y=j,Z=k}$ can be obtained as

$$P^{(CI)}_{Y=i,Z=k} = \sum_{i=1}^{I} P_{Y=j|X_D=i} \times P_{Z=k|X_D=i} \times P_{X_D=i}$$

and this value is always included in the corresponding interval identified by the uncertainty bounds:

$$P^{(low)}_{j,k} \leq P^{(CI)}_{Y=j,Z=k} \leq P^{(up)}_{j,k}, \qquad j=1,\ldots,J\ ;\ k=1,\ldots,K\,.$$

This property permits to view the approach to uncertainty as a kind of test on the CI assumption, the shorter will be interval the higher will be the trust on the CI to hold.

In the SM basic framework the probabilities $P_{Y=j|X_D=i}$ are estimated from A, the $P_{Z=k|X_D=i}$ are estimated from B, while the marginal distribution $P_{X_D=i}$ can be estimated indifferently on A or on B, assuming that both the samples, being representative samples of the same population, provide not different estimates of $P_{X_D=i}$.

In **StatMatch** the Fréchet bounds for $P_{Y=j,Z=k}$ ($j=1,\ldots,J$ and $k=1,\ldots,K$), conditioned or not on X_D, are calculated by `Frechet.bounds.cat` (already introduced in Section 3.1) which in input requires the estimated frequencies (absolute or relative) for the *X* variables (argument `tab.x`), the frequencies in the table $X_D \times Y$ (argument `tab.xy`) and the frequencies in the table $X_D \times Z$ (argument `tab.xz`). Note that data in input should have a coherent marginal distribution for the X_D variables, i.e. `tab.xy` and `tab.xz` should give the same marginal distribution provided by `tab.x`. The data used in the example are those of Appendix A.2; in particular, the bounds for the joint distribution of professional status (binary variable named `"c.pl030"`) and classes of net income (`"c.netI"`) are derived by conditioning on the joint distribution of age in classes (`"c.age"`) and gender (`"rb090"`). Note that the frequencies in the cells are computed by using the units' weights available in the data sources (variable `"wwA"` in *A* and `"wwB"` in *B*).

```
R> # exploring uncertainty on cells of c.pl030 vs. c.netI
R> # joint distr. of matching variables
R> #comparing joint distribution of the X_M variables in A and in B
R> t.xA <- xtabs(wwA~c.age+rb090, data=rec.A) ## wwA are units' weights
R> t.xB <- xtabs(wwB~c.age+rb090, data=don.B) ## wwB are units' weights
R> t.xx <- t.xA+t.xB  # expected distribution
R>
R> #computing tables needed by Frechet.bounds.cat
R> t.xy <- xtabs(wwA~c.age+rb090+work, data=rec.A)
R> t.xz <- xtabs(wwB~c.age+rb090+c.netI, data=don.B)
R> out.fb <- Frechet.bounds.cat(tab.x=t.xx, tab.xy=t.xy, tab.xz=t.xz,
R+                              print.f="data.frame")
R> out.fb
$bounds
          work   c.netI low.u       low.cx         CIA      up.cx
1      working   (-6,0]     0 0.0000000000 0.062064080 0.10660554
2  not working   (-6,0]     0 0.0134312975 0.057972760 0.12003684
3      working    (0,5]     0 0.0000000000 0.047448497 0.08051092
4  not working    (0,5]     0 0.0103308783 0.043393305 0.09084180
5      working   (5,10]     0 0.0000000000 0.065487967 0.10721825
```

```
6  not working   (5,10]     0 0.0335018576 0.075232143 0.14072011
7      working  (10,15]     0 0.0043439938 0.083386603 0.13234278
8  not working  (10,15]     0 0.0423161546 0.091272327 0.17031494
9      working  (15,20]     0 0.0307921002 0.105284154 0.15519978
10 not working  (15,20]     0 0.0342198421 0.084135471 0.15862753
11     working  (20,25]     0 0.0265260213 0.079654508 0.11262046
12 not working  (20,25]     0 0.0230109044 0.055976854 0.10910534
13     working  (25,30]     0 0.0034630255 0.042493435 0.06136383
14 not working  (25,30]     0 0.0060924885 0.024962882 0.06399329
15     working  (30,40]     0 0.0000000000 0.033215600 0.04838238
16 not working  (30,40]     0 0.0033326520 0.018499435 0.05171504
17     working  (40,50]     0 0.0000000000 0.008199235 0.01315724
18 not working  (40,50]     0 0.0001214154 0.005079418 0.01327865
19     working (50,200]     0 0.0000000000 0.010168702 0.01597007
20 not working (50,200]     0 0.0002712506 0.006072621 0.01624132
         up.u
1  0.11953297
2  0.11953297
3  0.09037855
4  0.09037855
5  0.14101622
6  0.14101622
7  0.17515499
8  0.17515499
9  0.18965562
10 0.18965562
11 0.13543995
12 0.13543995
13 0.06746229
14 0.06746229
15 0.05171910
16 0.05171910
17 0.01334022
18 0.01334022
19 0.01630008
20 0.01630008

$uncertainty
      av.u      av.cx    overall
0.10000000 0.07682461 0.11796992
```

The final component (`uncertainty`) of the output list provided by `Frechet.bounds.cat` summarizes the uncertainty by means of the average width of the unconditioned bounds, the average width of the bounds obtained by conditioning on X_D:

$$\bar{d} = \frac{1}{J \times K} \sum_{j,k} (\hat{P}_{j,k}^{(up)} - \hat{P}_{j,k}^{(low)})$$

and the overall uncertainty measured as suggested by Conti *et al.* (2012).

$$\hat{\Delta} = \sum_{i,j,k} \left(\hat{P}_{j,k}^{(up)} - \hat{P}_{j,k}^{(low)} \right) \times \hat{P}_{Y=j|X_D=i} \times \hat{P}_{Z=k|X_D=i} \times \hat{P}_{X_D=i}$$

When dealing with continuous variables, if the joint distribution of (X_M, Y, Z) is multivariate normal, the uncertainty bounds for the correlation coefficient ρ_{YZ} can be obtained as a direct consequence of the fact that the correlation matrix should be positive semidefinite. In the simple case of the trivariate normal distribution, the uncertainty on ρ_{YZ} due to the matching framework can be summarized as:

$$\rho_{XY}\rho_{XZ} - \sqrt{\left(1-\rho_{YX}^2\right)\left(1-\rho_{XZ}^2\right)} \leq \rho_{YZ} \leq \rho_{XY}\rho_{XZ} + \sqrt{\left(1-\rho_{YX}^2\right)\left(1-\rho_{XZ}^2\right)}$$

The higher are the correlations, between X and Y and between X and Z, the shorter will be the interval. It is worth noting that in such a case the CI solution, $\rho_{YZ}^{(CI)} = \rho_{XY}\rho_{XZ}$, is the central point of the interval.

In **StatMatch** it is possible to derive the bounds for ρ_{YZ} by calling the function `mixed.mtc` with argument `method="MS"`. The following example assumes multivariate normal distribution holding for joint distribution for age, gender (the matching variables), aggregated personal economic status (binary variable `"work"`, which plays the role of Y) and log-transformed personal net income (log of `"netIncome"`, is the Z variable).

```
R> don.B$log.netI <- log(ifelse(don.B$netIncome>0, don.B$netIncome, 0)
R+                       +1 )
R> X.mtc <- c("age","rb090")
R> mix.3 <- mixed.mtc(data.rec=rec.A, data.don=don.B, match.vars=X.mtc,
R+                    y.rec="work", z.don="log.netI",
R+                    method="MS")
input value for rho.yz is 0
low(rho.yz)= -0.8048
up(rho.yz)= 0.8667
The input value for rho.yz is admissible
```

6. Statistical matching of data from complex sample surveys

The SM techniques presented in the previous Sections implicitly or explicitly assume that *A* and *B* are i.i.d. samples from infinite population and the only source of variation is that of the model generating the data.

In sample surveys the inference is carried out in a different framework. The target population *U* is finite and consists of *N* units ($N < \infty$) and the set of the unknown values assumed by a generic variable *Y* for each unit in the population are fixed values and not outcomes of a random variable. The only randomization mechanism is the probability criterion used to select the sample *s* from *U*. The selection criterion assigns to each unit *U* of a non-null probability π_a ($0 < \pi_a \leq 1$) of being included into the sample. These *inclusion probabilities* π_a are the base of the inference.

Let assume that the objective of the inference is estimating the total amount of *Y* in *U*, $t_y = \sum_{a=1}^{N} y_a$, given a sample of *n* distinct units, selected according to a given generic

probabilistic criterion. It is possible to show that an estimate of t_y is provided by the Horvitz-Thompson estimator (HT) (cf. Sarndal, 1992):

$$\hat{t}_{\pi y} = \sum_{a=1}^{n} \frac{y_a}{\pi_a} = \sum_{a=1}^{n} d_a y_a$$

where $d_a = 1/\pi_a$ is the *design weight* or *base weight* or *direct weight* of the unit *a*. It is worth noting that the sum of the design weights of the sample units provides an estimate of the population size:

$$\hat{t}_{\pi 1} = \sum_{a=1}^{n} \frac{1}{\pi_a} = \sum_{a=1}^{n} d_a = \hat{N}$$

The HT estimator is an unbiased estimator of t_y with respect to the sampling design (*design unbiased*) whatever the probabilistic sampling design has been chosen.

For the inference purposes it is possible to ignore the probability distribution induced by the sampling design and treat the data as an i.i.d. sample (inferences are carried out in the model based framework) if the following conditions are verified:

i) *noninformative sampling design*: the sampling design does not depend on the $(y_1, y_2, \ldots, y_N)$ values, conditioned to the design variables (strata, etc.);

ii) the sampling design and the design variables are known and are not related to the parameters which are objective of the inference.

Ignoring the sampling design when these conditions do not hold may lead to unreliable results (see Section 13.6 in Sarndal *et al.*, 1992).

In general, when dealing with data from complex sample surveys it is common that the sampling design is not fully known or just a limited number of the design variable are available; moreover, in practice, when dealing with data sets originating from complex sample surveys it is common that:

- the observations in the data set are less than the planned sample size because of *unit nonresponse* (non-contacts, refusals, etc.) and of discarding of *ineligible units* (units no more belonging to the population but not deleted from the sampling frame before the selection of the sample);
- some units may present missing values; in other cases the missing values or the values identified as erroneous and therefore deleted have been substituted with imputed values;
- some of the observed values are affected by measurement errors (not detected by checks);
- the final weights w_a associated to the available units are the direct weights corrected to compensate for unit nonresponse, frame undercoverage and to reproduce known population totals concerning some important auxiliary variables;
- the design variables usually are partially available due to the risk of disclosure.

All these issues pose several problems when matching data of two complex surveys related to the same target population.

For our purposes, it will be assumed that each data source available in the SM framework is characterized by an initial sampling design (fully or partially known) and a set of final weights (w) (obtained by modifying the initial weights).

y_1	w_{A1}	x_{11}	$\cdots$	x_{p1}
y_2	w_{A2}	x_{12}	$\cdots$	x_{p2}
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
y_{m_A}	w_{Am_A}	x_{1m_A}	$\cdots$	x_{pm_A}

Survey A

Survey B

x_{11}	$\cdots$	x_{p1}	z_1	w_{B1}
x_{12}	$\cdots$	x_{p2}	z_2	w_{B2}
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
x_{1m_B}	$\cdots$	x_{pm_B}	z_{m_B}	w_{Bm_B}

In this case, the objective of SM can be the creation of a synthetic data set (micro) or the estimation of finite population parameters concerning the relationship between Y and Z (macro), e.g. correlation coefficient

$$\rho_{U,YZ} = \frac{\sum_{a=1}^{N}\left(y_a - \overline{y}_U\right)\left(z_a - \overline{z}_U\right)}{\sqrt{\sum_{a=1}^{N}\left(y_a - \overline{y}_U\right)^2 \sum_{a=1}^{N}\left(z_a - \overline{z}_U\right)^2}},$$

Or regression coefficient $B_{U,YZ}$, counts N_{jk} in the contingency table $Y \times Z$, or among X, Y and Z.

6.1 Naive micro approaches

A naive approach to SM of data from complex sample surveys consists in applying nonparametric micro methods (nearest neighbour distance, random or rank hot deck) without considering the sampling design nor the weights. Once obtained the synthetic dataset (recipient filled in with the missing variables) the successive statistical analyses are carried out by considering the sampling design underlying the recipient data set and the units' survey weights. In the following a simple example of constrained distance hot deck is reported by considering data introduced in the Appendix A.2.

```
R> summary(rec.A$wwA) # summary of weights in A
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  8.538  14.470  16.510  16.950  19.370  29.920

R> summary(don.B$wwB) # summary of weights in B
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
```

```
  6.149  10.580  11.890  12.280  13.950  21.550

R> # NND hot deck constrained
R> group.v <- c("rb090","db040")
R> out.nnd <- NND.hotdeck(data.rec=rec.A, data.don=don.B,
R+                        match.vars="age", don.class=group.v,
R+                        dist.fun="Manhattan",
R+                        constrained=TRUE, constr.alg="Hungarian")
R> fA.nnd.m <- create.fused(data.rec=rec.A, data.don=don.B,
R+                          mtc.ids=out.nnd$mtc.ids,
R+                          z.vars=c("netIncome","c.netI"))

R> # estimating average net income
R> weighted.mean(fA.nnd.m$netIncome, fA.nnd.m$wwA) # imputed in A
[1] 14940.63

R> weighted.mean(don.B$netIncome, don.B$wwB) # ref. estimate in B
[1] 15073.95
```

When imputing missing values in survey data, a way of taking into account the sampling design consists in forming the donation classes by using the design variables (stratification and/or clustering variables) jointly with the most relevant common variables (Andridge and Little, 2010). Unfortunately in SM this operation may be unfeasible because the design variables may be partly or not at all available. Moreover, the design variables used in one survey may be not available in the other one and vice versa, this typically happens when the data refer to surveys with different sampling designs.

Andridge and Little (2010) point out that when imputing the missing values with random hot deck, the selection of the donors can be carried out with probability proportional to weights associated to the donors (*weighted random hot deck*). This feature is implemented in the function `RANDwNDD.hotdeck`; the weighting variable should be specified via the `weight.don` argument, as shown in the following example.

```
R> group.v <- c("rb090","db040")
R> X.mtc <- "age"
R> rnd.2 <- RANDwNND.hotdeck(data.rec=rec.A, data.don=don.B,
R+                           match.vars=NULL, don.class=group.v,
R+                           weight.don="wwB")
```

The function `rankNDD.hotdeck` can use the units' weights (w_i) in A and in B when estimating of the percentage points of the empirical cumulative distribution function by means of the following expressions:

$$\hat{F}^{(A)}(x) = \frac{\sum_{i=1}^{n_A} w_i^{(A)} I\left(x_{A,i} \le x\right)}{\sum_{i=1}^{n_A} w_i^{(A)}}, \qquad \hat{F}^{(B)}(x) = \frac{\sum_{i=1}^{n_B} w_i^{(B)} I\left(x_{B,i} \le x\right)}{\sum_{i=1}^{n_B} w_i^{(B)}}$$

In this case, weights in both donor and recipient data set are used. The following code provides an example with percentage points of the empirical cumulative distribution function of the age, estimated region by region (variable `db040`).

```
R> rnk.w <- rankNND.hotdeck(data.rec=rec.A, data.don=don.B,
R+                          don.class="db040", var.rec="age",
R+                          var.don="age", weight.rec="wwA",
R+                          weight.don="wwB", constrained=TRUE,
R+                          constr.alg="Hungarian")
```

D'Orazio *et al.* (2012) compared several naive procedures. In general, when rank and random hot deck using the weights tend to perform quite well in terms of preservation in the synthetic data set of the marginal distribution of Z and of the joint distribution $X \times Z$. The nearest neighbour distance hot deck, provides good results only when constrained matching is used and a design variable (e.g. a stratification variable) is considered in forming donation classes.

6.2 Methods that account explicitly for the sampling weights

In literature there are few SM methods that explicitly take into account the sampling design and the corresponding sampling weights: Renssen's approach based on *weights' calibrations* (Renssen, 1998); Rubin's *file concatenation* (Rubin, 1986) and the approach based on the *empirical likelihood* proposed by (Wu, 2004). A comparison among these approaches can be found in D'Orazio *et al.* 2010 (see also D'Orazio, 2011a).

The Rubin's approach consist in concatenating the surveys $A \cup B$ and then computing a new inclusion probability for each unit in it:

$$\begin{aligned} \pi_{A\cup B,c} &= \pi_{A,c} + \pi_{B,c} - \pi_{A\cap B,c} \\ &\cong \pi_{A,c} + \pi_{B,c} \end{aligned} \qquad c \in A \cup B$$

The approximation holds when the chance of a unit of being included in two independent surveys carried out on the same finite population is close to zero (usually the samples are relatively small if compared to the population size). Unfortunately such an approach handles theoretic samples and does not account for reduction of sample size because of unit nonresponse. Moreover it may be quite difficult to derive the inclusion probabilities $\pi_{A\cup B,c}$ because it is required: (i) the knowledge of the sampling designs used to select *A* and *B* respectively; (ii) the design variables used to select *A* to be available in *A* and in *B*; and that (iii) the design variables used to select *B* to be available in *A* and in *B* (for major details on these problems see Ballin *et al.*, 2008). Once estimated the probabilities, it remains the problem of estimating the required parameters in presence of missing values because *Z* is missing in *A* and *Y* is missing in *B*.

The Renssen's approach is based on a series of steps of *calibration* of the survey weights in *A* and in *B*. In all the processing steps the two data sources are kept separate, and the final objective is that of estimating a parameter concerning the relationship between *Y* and *Z*.

Calibration is a technique very common in sample surveys for deriving new weights, as close as possible to the starting ones, which fulfil a series of constraints concerning totals for a set of auxiliary variables (for further details see Sarndal and Lundstrom, 2005). In particular, if w_k is the "starting" weights, the final calibrated weights $w_k^{(cal)}$ are derived as the solution of o minimization problem:

$$\min\left[\sum_{k\in r} D\left(w_k, w_k^{(cal)}\right)\right]$$

Subject to the following constraints (example with just one auxiliary variable *X*):

$$\sum_{k=1}^{m} w_c^{(cal)} x_k = \sum_{k=1}^{N} x_k \; ; \qquad \sum_{k=1}^{m} w_k^{(cal)} = N$$

where $D\left(w, w^{(cal)}\right)$ is a distance measure.

The Renssen's approach works well when dealing with categorical variables or in a mixed case in which the number of continuous variables is very limited.

Wu (2004) proposed the usage of *empirical likelihood* (EL) to combine data from multiple surveys (for major details on the EL see Chen and Sitter, 1999). In practice the technique is similar to the calibration, in fact it permits to derive new weights for units in *A* and for the units in *B* satisfying some constraints concerning the totals of *X* variables. These target totals can be: (i) known from external sources; (ii) estimated by combining estimates obtained separately from *A* and from *B* (Wu's *separate approach*) (similar to Renssen's approach); and (iii) unknown and not estimated, they are set to be equal as a further constraint in the optimization problem (Wu's *combined approach*). The Wu's approach is more flexible if compared to calibration (no negative weights) and in the combined case, it does not require to estimate the totals of the *X* variables. On the other hand it presents a major complexity in the presence of complex sampling designs; moreover it allows combining theoretic samples, but it does not handle reduced samples due to unit nonresponse.

In the following it will be shown in detail the Renssen's calibration approach, that is also implemented in the **StatMatch** package.

6.2.1 Statistical matching via weights calibration

The Renssen's procedure consists in estimating the two-way contingency table $Y \times Z$ starting from the data of two independent complex sample surveys carried out on the same finite population. It permits to exploit eventual auxiliary information represented by a third data source *C* in which (*X*, *Y*, *Z*) or simply (*Y*, *Z*) are available. It deals mainly with categorical variables. In order to estimate the contingency table $Y \times Z$, *linear probability models* are fitted by taking into account the survey weights in the estimation of the parameters. In this approach the data sets are maintained separate.

In the following it will be assumed that all the variables (X_D, Y, Z) are categorical, being X_D a complete or an incomplete crossing of the matching variables X_M. The first step of the procedure consists in calibrating the weights in A and in B such that the new weights

when applied to the set of the X_D variables allow to reproduce some known population totals. When the population totals of the X_D variables are unknown they are estimated by a weighted average of the totals estimated on the two surveys before the harmonization step:

$$\tilde{t}_{X_M} = \lambda \hat{t}^{(A)}_{X_M} + (1-\lambda)\hat{t}^{(B)}_{X_M}$$

usually $\lambda = n_A / (n_A + n_B)$ (Korn and Graubard, 1999, pp. 281-284), but the value λ ($0 \leq \lambda \leq 1$) be chosen according to different criteria.

The second step consists in estimating the two-way contingency table $Y \times Z$. In absence of auxiliary information it is estimated under the CI assumption by means of:

$$\hat{P}^{(CIA)}_{(Y=j,Z=k)} = \sum_{i=1}^{I} \hat{P}^{(A)}_{Y=j|X_D=i} \times \hat{P}^{(B)}_{Z=k|X_D=i} \times \hat{P}_{X_D=i}, \qquad i=1,\ldots,I;\ j=1,\ldots,J;\ K=1,\ldots,K.$$

In practice, $\hat{P}^{(A)}_{Y=j|X_D=i}$ is derived from A; $\hat{P}^{(B)}_{Z=k|X_D=i}$ is computed using data in B while $P_{X_D=i}$ can be estimated indifferently from A or B (the data set are harmonized with respect to the X_D distribution). Note that the relative frequencies are estimated by considering the units' weights ($w^{(1)}_k$) obtained after the initial harmonization step, i.e.,

$$\hat{P}_{X_D=i} = \frac{\sum_{k=1}^{n_A} w^{(1)}_k I(x_{Dk}=i)}{\sum_{k=1}^{n_A} w^{(1)}_k}$$

When a third data source C, with all the variables (X_M, Y, Z) or just (Y, Z), is available, the Renssen's approach permits to exploit it in estimating $Y \times Z$. Two alternative methods are available: (a) *incomplete two-way stratification*; and (b) *synthetic two-way stratification*. In practice, both the methods estimate the contingency table $Y \times Z$ from C after some further calibration steps.

The incomplete two-way stratification consists in (1) estimating the total amount of *Y* on *A* and the total of *Z* on *B*; (2) calibrating the weights of *C* to reflect these totals; and (3) finally use these new weights to estimate the contingency table $Y \times Z$ from *C*.

The synthetic two-way stratification, is less straightforward, in summary it starts from the contingency table $Y \times Z$ estimated under the CI assumption and then "corrects" it by considering the "distance" between it and the table $Y \times Z$ that can be estimated directly from *C* (for further details see Renssen, 1998).

In **StatMatch** the initial harmonization step can be performed by using the function `harmonize.x` that carries out weights' calibration (or post-stratification) by means of functions available in the R package **survey** (Lumley, 2012). When the population totals are already known then they have to be passed to `harmonize.x` via the argument `x.tot`; on the contrary, when unknown (`x.tot=NULL`) they are estimated by a weighted average of the totals from the two surveys as before shown. The following example shows how to harmonize the joint distribution of the gender and classes of age assuming that the "true" joint distribution of age and gender is not known.

```
R> install.packages("survey") # installs package survey
R> library("survey") # loads survey
R>
R> # joint distr. of gender vs. c.age in A
R> tt.A <- xtabs(wwA~rb090+c.age, data=rec.A)
R>
R> # joint distr. of gender vs. c.age in B
R> tt.B <- xtabs(wwB~rb090+c.age, data=don.B)
R> prop.table(tt.A)-prop.table(tt.B)
        c.age
rb090          [16,24]      (24,49]      (49,64]     (64,100]
  male    0.003661141  0.010995148 -0.009456418 -0.006383618
  female  0.000891681 -0.004970682  0.010772175 -0.005509426

R> # creates svydesign objects
R> svy.rec.A <- svydesign(~1, weights=~wwA, data=rec.A)
R> svy.don.B <- svydesign(~1, weights=~wwB, data=don.B)
R> #
R> # harmonizes wrt to joint distr. of gender vs. c.age
R> out.hz <- harmonize.x(svy.A=svy.rec.A, svy.B=svy.don.B,
R+                       form.x=~c.age:rb090-1)

R> summary(out.hz$weights.A) # new calibrated weights for A
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  8.647  14.390  16.570  16.950  19.030  31.470

R> summary(out.hz$weights.B) # new calibrated weights for B
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  6.279  10.540  11.840  12.280  13.910  22.400

R> # comparing joint distr. of Gender vs. Cl.of age in A and B
R> # after harmonization
R> tt.A <- xtabs(out.hz$weights.A~rb090+c.age, data=rec.A)
R> tt.B <- xtabs(out.hz$weights.B~rb090+c.age, data=don.B)
R> prop.table(tt.A) - prop.table(tt.B)
        c.age
rb090           [16,24]       (24,49]       (49,64]      (64,100]
  male    1.387779e-17 -2.775558e-17  0.000000e+00 -1.387779e-17
  female  0.000000e+00  2.775558e-17  0.000000e+00  0.000000e+00
```

The second step in the Renssen's procedure consists in estimating the two-way contingency table $Y \times Z$ and it is implemented in the function `comb.samples`. In absence of auxiliary information it is estimated under the CI assumption. The following example show how to estimate the distribution of professional status (variable `c.pl030`) vs. classes of net income (variable `c.netI`) starting from the two surveys after the harmonization step.

```
R> # estimating c.pl030 vs. c.netI under the CI assumption
R> out <- comb.samples(svy.A=out.hz$cal.A, svy.B=out.hz$cal.B,
R+                     svy.C=NULL, y.lab="work", z.lab="c.netI",
R+                     form.x=~c.age:rb090-1)
R> #
R> # table estimated under the CIA
```

```
R> addmargins(t(out$yz.CIA))
            working not working        Sum
(-6,0]    4203.9273   3929.4698  8133.3971
(0,5]     3212.7539   2941.5722  6154.3261
(5,10]    4436.4472   5108.0075  9544.4547
(10,15]   5648.5383   6199.2373 11847.7756
(15,20]   7129.6193   5716.1572 12845.7765
(20,25]   5391.3879   3802.7509  9194.1388
(25,30]   2877.6585   1696.1470  4573.8055
(30,40]   2249.5066   1256.9719  3506.4786
(40,50]    555.7829    345.2169   900.9998
(50,200]   688.8992    412.9481  1101.8473
Sum      36394.5210  31408.4790 67803.0000
```

When a third data source C is available the function `comb.samples` permits to estimate the contingency table $Y \times Z$ from C by means of the synthetic two-way stratification (argument `estimation="synthetic"`) or with the incomplete two-way stratification (argument `estimation="incomplete"`).

The incomplete two-way stratification estimates $Y \times Z$ from C by preserving the marginal distribution of Y and of Z estimated respectively from A and from B after the initial harmonization step; on the contrary, the joint distribution of the matching variables (which is the basis of the harmonization step) is not preserved. In this example data introduced in Appendix A.2 are used.

```
R> # incomplete two-way estimation
R> out.inc <- comb.samples(svy.A=out.hz$cal.A, svy.B=out.hz$cal.B,
R+                         svy.C=svy.aux.C, y.lab="work", z.lab="c.netI",
R+                         form.x=~c.age:rb090-1, estimation="incomplete")
R> #
R> # estimated table of work vs. c.net.I
R> addmargins(t(out.inc$yz.est))
            working not working        Sum
(-6,0]     318.3646   7815.0325  8133.3971
(0,5]     3155.6684   2998.6577  6154.3261
(5,10]    3960.8064   5583.6483  9544.4547
(10,15]   4736.0014   7111.7742 11847.7756
(15,20]   9302.3226   3543.4539 12845.7765
(20,25]   6318.9931   2875.1457  9194.1388
(25,30]   4011.6435    562.1620  4573.8055
(30,40]   2587.8739    918.6047  3506.4786
(40,50]    900.9998      0.0000   900.9998
(50,200]  1101.8473      0.0000  1101.8473
Sum      36394.5210  31408.4790 67803.0000
```

The synthetic two-way stratification (argument `estimation="synthetic"`) requires C to include the matching variables X_M too.

```
R> # synthetic two-way estimation
R> out.synt <- comb.samples(svy.A=out.hz$cal.A, svy.B=out.hz$cal.B,
R+                          svy.C=svy.aux.C, y.lab="work", z.lab="c.netI",
```

```
R+                          form.x=~c.age:rb090-1, estimation="synthetic")
R> #
R> # estimated table of work vs. c.net.I
R> addmargins(t(out.synt$yz.est))
            working not working        Sum
(-6,0]     351.6488   7781.7483  8133.3971
(0,5]     3610.2537   2544.0724  6154.3261
(5,10]    4052.7261   5491.7286  9544.4547
(10,15]   5384.8795   6462.8961 11847.7756
(15,20]   8542.0337   4303.7428 12845.7765
(20,25]   5971.5562   3222.5826  9194.1388
(25,30]   3781.3214    792.4840  4573.8055
(30,40]   2697.2545    809.2241  3506.4786
(40,50]    900.9998      0.0000   900.9998
(50,200]  1101.8473      0.0000  1101.8473
Sum      36394.5210  31408.4790 67803.0000
```

As for incomplete two-way stratification, the synthetic two-way stratification derives the table $Y \times Z$ from C by preserving the marginal distribution of Y and of Z estimated respectively from A and from B after the initial harmonization step; again, the joint distribution of the matching variables is not preserved.

The Renssen's approach has a macro objective (estimation of the contingency table $Y \times Z$) and for these purposes the *linear probability models* are fitted at unit level by taking into account the survey weights in the estimation of the regression parameters. The usage of the models enables to perform a regression imputation of the missing variables at micro level. In our case, such a procedure has some advantages; in fact, when all the variables (X_D, Y, Z) are categorical, the synthetic data set preserves the marginal distribution of the imputed Z variable and the joint distribution $X \times Z$. The main disadvantage is that the imputed value of Z for each unit in A corresponds to a vector of estimated probabilities: i.e. the probabilities that the given unit assumes one of the categories of the variable Z. Unfortunately linear probability models can provide negative or greater than one estimates for the probabilities (other well-known drawbacks of these models are heteroskedasticity and residuals not normally distributed). For these reasons, such predictions should be used carefully.

In `comb.samples` the estimated probabilities at unit level can be obtained by setting the argument `micro=TRUE`. In this case the function returns two additional data frames `Z.A` and `Y.B`. The first ones has the same rows as `svy.A` and the number of columns equals the number of categories of the Z variable (specified via `z.lab`). Each row provides the estimated probabilities for a unit of assuming a value in the various categories. The same happens for `Y.B` which presents the estimated probabilities of assuming a category of `y.lab` for each unit in B.

```
R> # predicting prob of c.netI in A under the CI assumption
R> out <- comb.samples(svy.A=out.hz$cal.A, svy.B=out.hz$cal.B,
R+                     svy.C=NULL, y.lab="work", z.lab="c.netI",
R+                     form.x=~c.age:rb090-1, micro=TRUE)

R> head(out$Z.A) # first 6 obs. of Z.A
         c.netI1    c.netI2    c.netI3   c.netI4   c.netI5    c.netI6
4547  0.02431737 0.03461536 0.07333853 0.1260644 0.2441140 0.22507260
9819  0.18449296 0.11651122 0.17192894 0.1828479 0.1536327 0.10775094
```

```
4461  0.01360657 0.02121963 0.08784363 0.1647151 0.2455691 0.14258653
10222 0.12862694 0.08280089 0.24704563 0.2624476 0.1531360 0.09119902
8228  0.18449296 0.11651122 0.17192894 0.1828479 0.1536327 0.10775094
3361  0.23596552 0.20079618 0.13191092 0.1593456 0.1707472 0.07521334
        c.netI7     c.netI8     c.netI9    c.netI10
4547  0.12213224 0.099898952 0.020044253 0.030402300
9819  0.04282870 0.024714585 0.009347124 0.005944963
4461  0.12907572 0.116517140 0.036077503 0.042789098
10222 0.01759877 0.011944017 0.003473668 0.001727448
8228  0.04282870 0.024714585 0.009347124 0.005944963
3361  0.02080676 0.005214446 0.000000000 0.000000000

R> # compare marginal distributions of Z
R> t.zA <- colSums(out$Z.A*out.hz$weights.A)
R> t.zB <- xtabs(out.hz$weights.B~don.B$c.netI)
R> prop.table(t.zA) - prop.table(t.zB)
don.B$c.netI
       (-6,0]         (0,5]        (5,10]       (10,15]       (15,20]
-8.326673e-17 -4.163336e-17  2.775558e-17 -2.775558e-17  1.110223e-16
      (20,25]       (25,30]       (30,40]       (40,50]      (50,200]
-2.775558e-17  2.775558e-17  3.469447e-17 -1.734723e-18  6.938894e-18
```

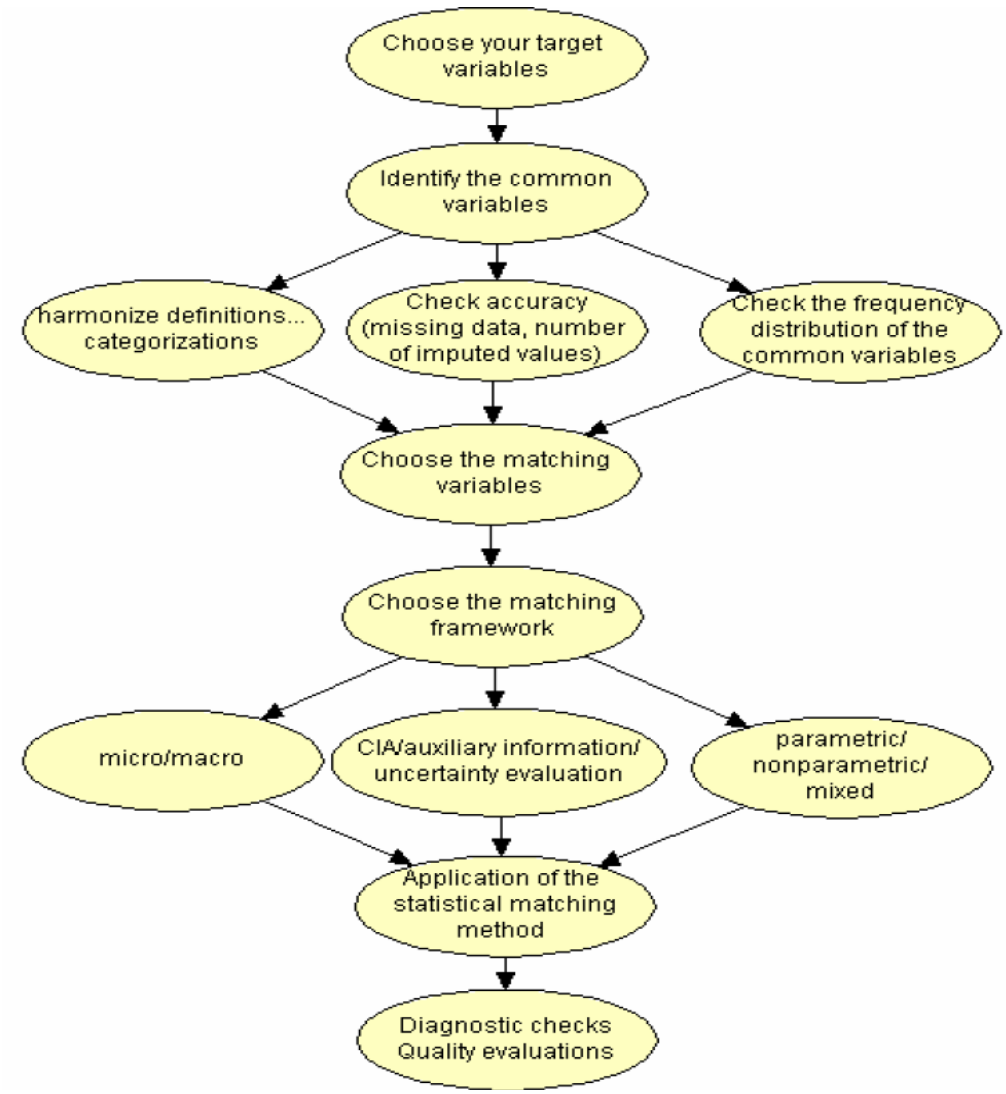
To derive the predicted category of each unit starting from the estimated probabilities a randomization scheme should be preferred to selecting the category with the highest estimated probability. For instance, it is possible to predict category in which a unit falls with probability proportional to the estimated probability (D'Orazio, 2012).

```
R> # check estimated probabilities
R> sum(out$Z.A<0) # negative est. prob.
[1] 0
R> sum(out$Z.A>1) # est. prob. >1
[1] 0

R> # predicting class of netIncome in A
R> # randomized prediction with prob proportional to estimated prob.
R> pred.zA <- apply(out$Z.A,1,sample,x=1:ncol(out$Z.A), size=1,replace=F)
R> rec.A$c.netI <- factor(pred.zA, levels=1:nlevels(don.B$c.netI),
R+                        labels=as.character(levels(don.B$c.netI)),
R+                        ordered=T)
R> #
R> # comparing marginal distributions of Z
R> t.zA <- xtabs(out.hz$weights.A~rec.A$c.netI)
R> t.zB <- xtabs(out.hz$weights.B~don.B$c.netI) # origin in B
R> prop.table(t.zA) - prop.table(t.zB)
rec.A$c.netI
       (-6,0]         (0,5]        (5,10]       (10,15]       (15,20]
 0.0016204734 -0.0047532489  0.0029080482  0.0046909159  0.0149523571
      (20,25]       (25,30]       (30,40]       (40,50]      (50,200]
-0.0157371610 -0.0026500936 -0.0024265175 -0.0005240009  0.0019192274
```

7. Practical problems in Statistical Matching

Before applying SM methods to integrate two or more data sources some decisions and some pre-processing steps are required (Scanu, 2008). The following diagram, provides a summary picture of the various steps in a SM application (Chapter 3 in ESSnet Report of WP2 2009).



It is worth noting that most of the actions are strictly related each other. For instance the choice of the marching variables is strictly related to the matching framework.

7.1 Identification of the common variables

The identification of all the common variables X shared by A and B may require harmonization of different definitions and/or classifications. If two similar variables cannot be harmonized they have to be discarded. The common variables should not present missing

values and the observed values should be accurate.; moreover, they are expected to share the same marginal/joint distribution, if *A* and *B* are representative samples of the same population. The comparison of the marginal distributions of the common variables requires the application of statistical tests (Chi-Square, Kolmogorov-Smirnov, etc.). When dealing with data from complex sample surveys *ad hoc* test accounting for complex sampling designs should be considered (for modified Chi-Square tests cf. Sarndal *et al.*, 1992, pp. 500-513). Unfortunately just a limited number of modified test it is available and their application requires additional information (sampling errors or design effects, etc.) that may not be available. For these reasons sometimes it is preferred an "empirical approach" which consists in comparing the marginal distributions estimated from the two surveys by means of *similarity/dissimilarity measures*.

The *dissimilarity index* or *total variation distance* between distributions is:

$$\Delta_{AB} = \frac{1}{2}\sum\nolimits_{c=1}^{C}\left|\hat{P}_{Ac} - \hat{P}_{Bc}\right| = \frac{1}{2}\sum\nolimits_{s=1}^{2}\sum\nolimits_{c=1}^{C}\left|\hat{P}_{sc} - \hat{P}_{+c}\right|$$

It ranges among 0 (minimum dissimilarity, the distributions are equal) and 1, and can be interpreted as the smallest fraction of units that would need to be re-classified in order to make the distributions equal. When the index is used to assess the closeness of an observed distribution to the expected values under a given hypothesis (model) then, following Agresti (2002, pp. 329-330), $\Delta < 0.02$ or 0.03, denotes that "*the sample data follow the model pattern quite closely, even though the model is not perfect*".

It is worth noting that, when comparing estimated marginal distributions of the same variable but derived from independent data sources, the expected distribution of the variable can be obtained as a weighted average of the two estimates obtained separately from the two sample surveys:

$$\hat{P}_{+c} = \frac{n_A\hat{P}_{Ac} + n_B\hat{P}_{Bc}}{n_A + n_B} = \lambda_A\hat{P}_{Ac} + \left(1-\lambda_A\right)\hat{P}_{Bc},$$

where usually $\lambda_A = n_A/\left(n_A + n_B\right)$. Thus, when comparing both the distributions with this expected one, following the Agresti's rule of thumb, that they can be considered coherent if $\Delta_{AB} \le 0.06$.

The total variation distance can be written as:

$$\Delta_{AB} = 1 - OV_{AB} = 1 - \sum\nolimits_{c=1}^{C}\min\left(\hat{P}_{Ac}, \hat{P}_{Bc}\right)$$

where "OV" stays for the *overlap* between the two distributions ($0 \le OV_{AB} \le 1$). The overlap is a similarity measure that equals 1 when the two distributions are equal.

A distance between the two distributions can be computed by means of the *Hellinger's Distance*:

$$d_{H,AB} = \sqrt{1 - BC_{AB}} = \sqrt{1 - \sum\nolimits_{c=1}^{C}\sqrt{\hat{P}_{Ac}\times\hat{P}_{Bc}}}$$

where BC is the *Bhattacharyya coefficient* ($0 \le BC \le 1$). The Hellinger's distance satisfies the properties of a distance ($0 \le d_H \le 1$, symmetry, triangle inequality); unfortunately it is not possible to determine a threshold of acceptable values of the distance according to which the two distributions can be said close. However, it is possible to show that:

$$d_H^2 \le \Delta \le d_H \sqrt{2}$$

According to this inequality, it comes out that admitting values such that $\Delta \le 0.06$ means admitting an Hellinger distance $d_H \le 0.042$ (a rule of thumb often recurring in literature considers two distributions close if the Hellinger's distance is not greater than 0.05).

In **StatMatch** the comparison between marginal or joint distributions of categorical variables can be done via the function `comp.prop` which returns some similarity/dissimilarity measures between distributions (dissimilarity index, the overlap between distributions, the Bhattacharyya's coefficient and the Hellinger's distance) and performs the Chi-square test too. A crucial argument in `comp.prop` is `ref`: when `ref=TRUE`, it means that the second distribution (argument `p2`) is the reference one and we look how close is the first distribution (argument `p1`) w.r.t to second one; on the contrary, when `ref=FALSE` the two distributions are just compared (and the reference one is obtained as a weighted average between the two; `p.exp` in the output):

```
R> # comparing joint distr. of Gender vs. Cl.of age in A and B
R> # no one is the reference distribution
R> tt.A <- xtabs(wwA~rb090+c.age, data=rec.A)
R> tt.B <- xtabs(wwB~rb090+c.age, data=don.B)
R> comp.prop(p1=tt.A, p2=tt.B, n1=nrow(rec.A), n2=nrow(don.B), ref=FALSE)
$meas
       tvd    overlap      Bhatt       Hell
0.02632014 0.97367986 0.99956825 0.02077856

$chi.sq
   Pearson         df      q0.05   delta.h0
 8.0082627  7.0000000 14.0671404  0.5692886

$p.exp
         c.age
rb090       [16,24]    (24,49]    (49,64]   (64,100]
  male   0.07010041 0.22316357 0.11129434 0.07671609
  female 0.06578194 0.22773205 0.11842264 0.10678897
```

Comparing the marginal distributions of continuous variables poses major problems. Descriptive statistics (minimum, maximum, average, standard deviation, coefficient of variation, percentiles) and graphical analysis can be of help. In alternative, the continuous variables can be categorized and the tools previously introduced can be used.

7.2 Choice of the matching variables

The set of common variables shared by A and B may be quite large and usually not all of them are used in SM. A subset of them, the *matching variables* X_M ($X_M \subseteq X$), should be selected through opportune statistical methods (descriptive, inferential, etc.), bearing in mind the matching framework (see Table 1). Cohen (1991) points out that the choice should be carried out in a "multivariate sense" in order to identify the subset X_M connected at the same time with Y and Z; unfortunately this would require the availability of an auxiliary data source in which all the variables (X, Y, Z) are observed.

In the basic SM framework the data in A permit to explore the relationship between Y and X, while the relationship between Z and X can be investigated in the file B. Then, the results of the two separate analyses have to be combined in some manner. Usually the subset of the matching variables is obtained as $X_M = X_Y \cup X_Z$, being X_Y ($X_Y \subseteq X$) the subset of the common variables that better explains Y, while X_Z is the subset of the common variables that better explain Z ($X_Z \subseteq X$). Unfortunately, such a procedure may result in too many matching variables, which consequently may increase the complexity of the problem and potentially affect negatively the results of SM. In particular, in the micro approach this may introduce additional undesired variability and bias, as far as the joint distribution of X_M and Z is concerned. For this reason in most of the cases the set of the matching variables is obtained as a compromise:

$$X_Y \cap X_Z \subseteq X_M \subseteq X_Y \cup X_Z$$

Subject matter experts can be of help in finding a good balance between the two extremes.

The simplest procedure to identify X_Y consists in computing pairwise correlation/association measures between Y and all the available predictors X. The same procedure is applied in B to derive X_Z. The measures to apply depends on the nature on the variables being considered:

a) <u>*Y* continuous or categorical ordered, *X* continuous or categorical ordered</u>.
It can be used the *Spearman correlation coefficient* (correlation coefficient computed on the ranks of *Y* (*s*) and *X* (*r*)):

$$\rho_S = \frac{\sum_{i=1}^{n_A} (r_i - \bar{r})(s_i - \bar{s})}{\sqrt{\sum_{i=1}^{n_A} (r_i - \bar{r})^2 \sum_{i=1}^{n_A} (s_i - \bar{s})^2}}$$

The Spearman correlation can detect nonlinear monotonic relationship. Moreover $\rho_S^2 = R_{sr}^2$ is the coefficient of determination of the linear regression of rank(*Y*) vs. rank(*X*). In this sense, Harrell (2001) suggests considering the adjusted version of R^2. When non-monotonic relationship is supposed to exist among variables (e.g. "U" shaped relationship) Harrell suggests considering R_{adj}^2 of the linear regression:

rank(Y) vs. rank(X) + [rank(X)]^2

b) Y continuous or categorical ordered, X categorical nominal.
The ranks of Y values are considered, while X is substituted by the corresponding $I-1$ dummy variables (I are the categories of X). As in the previous case it can be considered the coefficient of determination of the regression

rank(Y) vs. dummies(X)

In this case the coefficient of determination corresponds to the Eta-squared (η^2) measure of effect size used in ANOVAhttp://en.wikiversity.org/wiki/ANOVA but computed on the rank of the response variable:

$$\eta_s^2 = \frac{\sum_{i=1}^{I} n_i \left(\overline{s}_i - \overline{s}\right)^2}{\sum_{i=1}^{I}\sum_{a=1}^{n_i}\left(s_{ia} - \overline{s}\right)^2}$$

such a measure is strictly related to the Kruskal-Wallis test statistic.

c) Y categorical nominal, X categorical nominal (or categorical ordered).
The contingency table of X vs. Y is derived and Chi-Squared association measures can be computed:

$$\chi^2 = \sum_{i=1}^{I}\sum_{j=1}^{J}\frac{\left(p_{ij} - p_{i+}p_{+j}\right)^2}{p_{i+}p_{+j}}$$

being p_{ij} the estimated relative frequency. Sometimes, the *Cramer's V* it is considered:

$$V = \sqrt{\frac{\chi^2}{\min\left[(G-1),(H-1)\right]}}$$

it is a relative measure of association between two variables ($0 \le V \le 1$) and expresses the association as a percentage of the maximum possible variation. Unfortunately the more unequal are the marginal distribution of the variables being considered the more V will be less than 1.

It is possible to derive measures similar to the R^2 by reasoning in terms of variance. In particular, it should be considered a measure of the *proportional reduction of the variance* of Y when passing from the marginal distribution to the conditional distribution given X (cf. Agresti, 2002, p. 56):

$$\frac{V(Y) - E\left[V\left(Y|X\right)\right]}{V(Y)}$$

where $E\left[V\left(Y\middle|X\right)\right]=\sum_{i=1}^{I}p_{i+}V\left(Y\middle|i\right)$.

Unfortunately with categorical variables there is not a general accepted definition of variance. If variance is measured in terms of *entropy*:

$$V\left(Y\right)=-\sum_{j=1}^{J}p_{+j}\log p_{+j}$$

and the proportional reduction of variance formula gives the Theil's *uncertainty coefficient* (cf. Agresti, 2002, p. 56):

$$U_{YX}=\frac{-\sum_{j=1}^{J}p_{+j}\log p_{+j}+\sum_{i=1}^{I}\sum_{j=1}^{J}p_{ij}\log\left(p_{ij}/p_{i+}\right)}{-\sum_{j=1}^{J}p_{+j}\log p_{+j}}$$

It provides the relative reduction of uncertainty when predicting *Y* using the information provided by *X*. $U_{YX}=0$ denotes that *X* is not of help in predicting *Y*.

When *concentration* is considered:

$$V\left(Y\right)=1-\sum_{j=1}^{J}p_{+j}^{2}$$

and the Goodman & Kruskal *concentration coefficient* comes out:

$$\tau_{YX}=\frac{\sum_{i=1}^{I}\sum_{j=1}^{J}p_{ij}^{2}/p_{i+}-\sum_{j=1}^{J}p_{+j}^{2}}{1-\sum_{j=1}^{J}p_{+j}^{2}}$$

It can be interpreted as the proportional reduction in the probability of an incorrect guess when predicting *Y* using *X* ($\tau=0$ when the variables are independent).

If one looks at *classification errors*:

$$V\left(Y\right)=1-\max_{j}\left(p_{+j}\right)$$

and the Goodman & Kruskal λ association measure results:

$$\lambda_{YX}=\frac{\sum_{i=1}^{I}\max_{j}\left(p_{ij}\right)-\max_{j}\left(p_{+j}\right)}{1-\max_{j}\left(p_{+j}\right)}$$

This measure is interpreted as the proportional reduction in error when predicting *Y* based on *X*. This is easier to interpret than *U* and τ, but $\lambda_{YX}=0$ does not imply independence (the opposite is true), it just denotes the absence of predictive association (the knowledge of the *X* category is not of help in predicting the *Y* variable) (cf. Bishop et al. 1995).

In R the function `spearman2` in the package **Hmisc** (Harrell *et al.*, 2012) computes automatically the adjusted R^2 for each couple response-predictor (when response is a continuous or categorical ordered variable).

```
R> intersect(names(rec.A), names(don.B)) # variables with same name
[1] "hsize"  "hsize6" "db040"  "age"    "c.age"  "rb090"  "pb220a" "rb050"
[9] "c.netI"
R> install.packages("Hmisc") # installs package Hmisc
R> require("Hmisc") # loads package Hmisc
R> # analyses on B
R> spearman2(netIncome~db040+hsize+age+rb090+pb220a, data=don.B)

Spearman rho^2    Response variable:netIncome

        rho2      F df1  df2      P Adjusted rho2    n
db040  0.003   2.20   8 5513 0.0243         0.002 5522
hsize  0.030 170.97   1 5520 0.0000         0.030 5522
age    0.032 184.98   1 5520 0.0000         0.032 5522
rb090  0.147 952.42   1 5520 0.0000         0.147 5522
pb220a 0.018  50.98   2 5519 0.0000         0.018 5522
```

When dealing with couples of categorical variables, the pairwise association measures can be derived by using the function `pw.assoc` in the package **StatMatch**, as shown in the following example.

```
> # analyses on A
> pw.assoc(pl030~db040+hsize6+c.age+rb090+pb220a, data=rec.A)
$V
 pl030.db040 pl030.hsize6  pl030.c.age  pl030.rb090 pl030.pb220a
  0.07369617   0.19172123   0.52701354   0.43451872   0.11761739

$lambda
 pl030.db040 pl030.hsize6  pl030.c.age  pl030.rb090 pl030.pb220a
  0.00000000   0.05476951   0.27339115   0.00000000   0.00000000

$tau
 pl030.db040 pl030.hsize6  pl030.c.age  pl030.rb090 pl030.pb220a
 0.005804228  0.053874437  0.245431041  0.054777396  0.004970513

$U
 pl030.db040 pl030.hsize6  pl030.c.age  pl030.rb090 pl030.pb220a
 0.010376238  0.065400904  0.272715553  0.073490286  0.009215848
```

More sophisticated methods can be used for the selection of the best predictors. For instance, when fitting a linear regression model it is possible to use automatic procedures (backward, forward, stepwise) to reduce the predictors. Such procedures should be used with a certain care; it would be preferable to use procedure based on the *residual Chi-square* or the *Akaike's information criterion* (AIC) (cf. Harrell, 2001). When all the predictors are continuous *Least Angle Regression* procedures can be used (Efron *et al.*, 2004). Recently,

similar procedures have been developed for linear regression model with ordinal predictors (Gertheiss, 2011) for the *Generalised Linear Mixed Models* (GLMM) (Groll & Tutz, 2011).

When carrying out regression analysis on data from complex sample surveys it would be preferable to compare the results obtained under the i.i.d. assumption with those obtained by applying regression models that take into account for the sampling design: (i) explicitly, i.e. sampling design and weights are considered explicitly by ad hoc methods (e.g. *design weighted least squares*; or (ii) indirectly, i.e. the design variables are included into the model as explanatory variables.

In complex cases (categorical response and/or mixed type predictors), fitting nonparametric regression models can be of help. *Classification And Regression Trees* (CART; Breiman et al. 1984) can detect nonlinear relationship among response and the predictors. Unfortunately, basic CART suffer of some well-known drawbacks: selection bias (tend to prefer predictors with many possible splits) and collinearity. For this reason it would be better to resort to random Forests procedures (Breiman, 2001). Random forest procedure, present the further advantage of providing measures of predictors' importance (to be used carefully).

When there are too many common variables, before searching the best predictors it would be preferable to discard redundant predictors with *redundancy analysis* (cf. Harrell 2012) or simple explorative procedures based on *variable clustering* (Sarle, 1990; Harrell 2012).

An approach for selecting matching variables that goes in the direction of reasoning in a "multivariate sense" relies on the evaluation of the uncertainty due to the matching framework. In particular, the idea is that of searching for the subset of common variables more effective in reducing this uncertainty. In particular, when all the variables (X, Y, Z) are categorical, assuming that X_D corresponds to the complete crossing of some of the X variables, it is possible to recall results introduced in Section 5:

$$P_{j,k}^{(low)} \leq P_{Y=j,Z=k} \leq P_{j,k}^{(up)},$$

with

$$P_{j,k}^{(low)} = \sum_{i=1}^{I} P_{X_D=i} \times \max\left\{0;\, P_{Y=j|X_D=i} + P_{Z=k|X_D=i} - 1\right\}$$

$$P_{j,k}^{(up)} = \sum_{i=1}^{I} P_{X_D=i} \times \min\left\{P_{Y=j|X_D=i};\, P_{Z=k|X_D=i}\right\}$$

for $j = 1, \ldots, J$ and $k = 1, \ldots, K$; being J and K the categories of Y and Z respectively.

The function `Fbwidths.by.x` in **StatMatch** estimates $(P_{j,k}^{(low)}, P_{j,k}^{(up)})$ for each cell in the contingency table $Y \times Z$ in correspondence of all the possible combinations of the subsets of the X variables; then the reduction of uncertainty is measured according to the proposal of Conti *et al.* (2012):

$$\hat{\Delta} = \sum_{i,j,k} \left(\hat{P}_{j,k}^{(up)} - \hat{P}_{j,k}^{(low)}\right) \times \hat{P}_{Y=j|X_D=i} \times \hat{P}_{Z=k|X_D=i} \times \hat{P}_{X_D=i}$$

or, naively, by considering the average widths of the intervals:

$$\bar{d} = \frac{1}{J \times K} \sum_{j,k} (\hat{P}_{j,k}^{(up)} - \hat{P}_{j,k}^{(low)})$$

```
R> # estimate the contingency tables needed by Fbwidths.by.x
R> xx <- xtabs(~db040+hsize6+c.age+rb090+pb220a, data=rec.A)
R> xy <- xtabs(~db040+hsize6+c.age+rb090+pb220a+work, data=rec.A)
R> xz <- xtabs(~db040+hsize6+c.age+rb090+pb220a+c.netI, data=don.B)

R> out.fbw <-  Fbwidths.by.x(tab.x=xx, tab.xy=xy, tab.xz=xz)

R> # results sorted according to overall uncertainty
R> sort.ov.unc <- out.fbw$sum.unc[order(out.fbw$sum.unc$ov.unc),]
R> head(sort.ov.unc) # best 6 subsets of Xs

                                x.vars x.cells x.freq0   av.width    ov.unc
|c.age+rb090                         2       8       0 0.07738444 0.1204430
|c.age                               1       4       0 0.08439346 0.1213526
|hsize6+c.age                        2      24       0 0.08282457 0.1236898
|db040+hsize6+c.age+rb090+pb220a     5    1296     721 0.05775534 0.1238597
|c.age+rb090+pb220a                  3      24       1 0.07642623 0.1246536
|c.age+pb220a                        2      12       0 0.08432518 0.1257270
```

In practice, with these example data, the smallest value of the overall uncertainty measured by means of $\hat{\Delta}$ is achieved when (`c.age`) and gender (`rb090`) are considered, thus confirming the finding of the previous analyses.

7.3 Assessing accuracy of results of statistical matching

Evaluating the results of statistical matching is very difficult because:

(i) the objective of the inference is to study the relationship of phenomena not jointly observed, unless an additional auxiliary data source is available.

(ii) The statistical matching can provide different outputs: a synthetic data set in the micro case or estimates of parameters (e.g. a correlation coefficient, a regression coefficient, probabilities in a contingency table, etc.) in the macro case.

(iii) The available data sources may have different quality "levels" (sampling design, sample size, data processing steps, etc.);

The issue (i) is crucial, it is the major source of uncertainty concerning the matching results. In practice the application of statistical matching requires necessarily to fill in the lack of information concerning the relationship between *Y* and *Z* by making some assumptions (e.g. the conditional independence of the target variables given the matching variables) or using additional auxiliary information (an external estimate of the interest parameters or an additional data source). In this sense, the result of the statistical matching will necessarily reflect the underlying assumptions/information being used; results of a matching application based on the CI assumption will reflect it, unless some additional noise or bias is introduced

in the matching process. Clearly the results will be unreliable if CI is not holding. On the other hand, if CI assumption is avoided by using some kind of auxiliary information (past surveys or additional data sources) it is expected that the result of the SM will reflect such input (unless some additional noise is introduced in the matching process). Again, if the input information is not reliable, the results of SM will be unreliable.

In this setting a researcher has already some expectations concerning the SM outputs, he however has to apply some additional control actions to ascertain whether the chosen matching method has been applied correctly avoiding the introduction of additional noise or bias.

The approach consisting in the evaluation of the uncertainty due to the matching framework can be of help in assessing the reliability of the results of the SM. In the simple case of three continuous variables (*X*,*Y*,*Z*) following the trivariate normal distribution (*X* and *Y* available in *A* and X and *Z* observed in *B*) it is known that:

$$\rho_{XY}\rho_{XZ} - \sqrt{\left(1-\rho_{XY}^2\right)\left(1-\rho_{XZ}^2\right)} \leq \rho_{YZ} \leq \rho_{XY}\rho_{XZ} + \sqrt{\left(1-\rho_{XY}^2\right)\left(1-\rho_{XZ}^2\right)}$$

Obviously, the shorter is the interval, less uncertainty on ρ_{YZ} there will be. It is worth noting that the midpoint of the interval, $\rho_{XY}\rho_{XZ}$, is the expected value under the CI assumption: $\rho_{YZ}^{(\mathrm{CI})} = \rho_{XY}\rho_{XZ}$. In this terms, evaluating the uncertainty can be considered a kind of test on the CI assumption; the shorter is the uncertainty interval the higher will be the trust in the CI.

In the case of categorical *X*, *Y* and *Z* variables, the uncertainty can assessed by considering the Fréchet bounds (see Section 5 or 7.2):

$$P_{j,k}^{(\mathrm{low})} \leq P_{Y=j,Z=k} \leq P_{j,k}^{(\mathrm{up})}$$

In this case the CI solution is still included in the interval:

$$P_{j,k}^{(\mathrm{low})} \leq P_{j,k}^{(\mathrm{CI})} \leq P_{j,k}^{(\mathrm{up})}$$

but it is not the midpoint.

Uncertainty offers a way to evaluate accuracy of the results when the objective of the SM is macro. In the micro case, things are more complex and the only possibility consists in assessing how the synthetic data set can be considered a representative sample of the target population. Obviously a judgment of the reliability of the results can be derived by the validity of the underlying assumptions and some indicators concerning the integration process.

The representativeness of the synthetic data set can be evaluated only indirectly by introducing some checks on the variables used in the SM application. As suggested by Rässler's (2002) one looks at the "validity" of the SM procedure by analysing how the synthetic data set:

a) preserves the marginal distribution of the imputed variable (the reference is the marginal distribution estimated from the donor data set);

b) preserves the joint distribution of the imputed variable with the matching variables (the reference is the joint distribution estimated from the donor data set).

In order to compare marginal or joint distributions of the variables in the synthetic data set with respect to the donor, it is possible to use tests and descriptive measures introduced in Section 7.1. In particular when dealing with categorical variables the function `comp.prop` can be applied; in such a case we compare the marginal distribution estimated on the synesthetic data set with the corresponding one in the donor data set that represents the reference, as shown in the following example.

```
R> # comparing distributions in the synthetic wrt to the donor
R> # constrained distance hot deck
R> out.nnd.c <- NND.hotdeck(data.rec=rec.A, data.don=don.B,
R+                          match.vars="age", don.class=group.v,
R+                          dist.fun="Manhattan",
R+                          constrained=TRUE, constr.alg="Hungarian")
Warning: The  Manhattan  distance is being used
All the categorical matching variables in rec and don
 data.frames, if present are recoded into dummies
R> fA.nnd.c <- create.fused(data.rec=rec.A, data.don=don.B,
R+                          mtc.ids=out.nnd.c$mtc.ids,
R+                          z.vars=c("netIncome","c.netI"))
R>
R> # marginal tables in donor (reference)
R> t.Z <- xtabs(~c.netI, data=don.B) # Z variable
R> t.xZ <- xtabs(~rb090+db040+c.age+c.netI, data=don.B) # X vs Z
R>
R> # marginal tables in synthetic (reference)
R> t.Z.sc <- xtabs(~c.netI, data=fA.nnd.c) # Z variable
R> t.xZ.sc <- xtabs(~rb090+db040+c.age+c.netI, data=fA.nnd.c) # X vs Z
R>
R> # comparing distributions
R> out.1c <- comp.prop(p1=t.Z, p2=t.Z.sc, n1=nrow(fA.nnd.c), ref=TRUE)
R> out.1c$meas
       tvd    overlap      Bhatt       Hell
0.02838718 0.97161282 0.99944975 0.02345732
R> out.2c <- comp.prop(p1=t.xZ, p2=t.xZ.sc, n1=nrow(fA.nnd.c), ref=TRUE)
R> out.2c$meas
      tvd   overlap     Bhatt      Hell
0.1832769 0.8167231 0.9560245 0.2097034
```

Appendix A
R code for generating data used in the examples

A.1 Data sets generated from a multivariate normal distribution

In R to generate data from a given multivariate Normal distribution it is possible to resort to the package **mvtnorm** (Genz, 2013)

```
R> install.packages("mvtnorm") # to install the package
R> library(mvtnorm)  # to load the package
R>
R> # define a 3x3 correlation matrix
R> cc <- diag(1,3)
R> cc[1,2] <- cc[2,1] <- 0.55
R> cc[1,3] <- cc[3,1] <- 0.85
R> cc[2,3] <- cc[3,2] <- 0.80
R> dimnames(cc) <- list(c("x","y","z"), c("x","y","z"))
R> cc
     x    y    z
x 1.00 0.55 0.85
y 0.55 1.00 0.80
z 0.85 0.80 1.00
R>
R> #generate A with of n_A=50 obs
R>
R> data.A <- rmvnorm(50, mean=c(0,0,0), sigma=cc)
R> colnames(data.A) <- c("x","y","z")
R> data.A <- data.A[,-3] # drop z
R> head(data.A)
                x          y
[1,] -1.47011062  0.2231917
[2,]  0.22957161  2.6339102
[3,]  0.04839709 -0.9714869
[4,]  0.84314453  1.4781465
[5,]  1.71006937  1.9395058
[6,]  0.01294460  0.1524463
```

A.2 Data sets derived from artificial EU-SILC data

Artificial data derived resembling a subset of the variables of the EU-SILC (Survey on Income and Living Conditions) can be generated by using the data set `eusilcS` (artificial data generated from real Austrian EU-SILC survey ; `eusilcS` help pages for details) provided by the R package **simPopulation** (Alfons and Kraft, 2012). The following R code shows how the A and B data set have been derived starting from `eusilcS`.

```
R> install.packages("simPopulation")  # run just once to install package
R> library("simPopulation")  # loads package simPopulation
R> data("eusilcS")
R> str(eusilcS)

'data.frame':    11725 obs. of  18 variables:
 $ db030    : int  1 1 2 3 4 4 4 5 5 5 ...
 $ hsize    : int  2 2 1 1 3 3 3 5 5 5 ...
 $ db040    : Factor w/ 9 levels "Burgenland",..: 4 4 7 5 7 7 7 4 4 4 ...
 $ age      : int  72 66 56 67 70 46 37 41 35 9 ...
 $ rb090    : Factor w/ 2 levels "male","female": 1 2 2 2 2 1 1 1 2 2 ...
 $ pl030    : Factor w/ 7 levels "1","2","3","4",..: 5 5 2 5 5 3 1 1 3  ...
 $ pb220a   : Factor w/ 3 levels "AT","EU","Other": 1 1 1 1 1 1 3 1 1 ...
 $ netIncome: num  22675 16999 19274 13319 14366 ...
 $ py010n   : num  0 0 19274 0 0 ...
 $ py050n   : num  0 0 0 0 0 ...
 $ py090n   : num  0 0 0 0 0 ...
 $ py100n   : num  22675 0 0 13319 14366 ...
 $ py110n   : num  0 0 0 0 0 0 0 0 0 NA ...
 $ py120n   : num  0 0 0 0 0 0 0 0 0 NA ...
 $ py130n   : num  0 16999 0 0 0 ...
 $ py140n   : num  0 0 0 0 0 0 0 0 0 NA ...
 $ db090    : num  7.82 7.82 8.79 8.11 7.51 ...
 $ rb050    : num  7.82 7.82 8.79 8.11 7.51 ...
```

Before using data for our purposes, some manipulations are needed to discard units not relevant (people with `age<16`, whose income and personal economic status are missing), to categorize some other variables, etc.

```
R> # discard units with age<=16
R> silc.16 <- subset(eusilcS, age>15)

R> # categorize age
R> silc.16$c.age <- cut(silc.16$age, c(16,24,49,64,100),
R+                      include.lowest=TRUE)

R> # truncate hsize
R> aa <- as.numeric(silc.16$hsize)
R> aa[aa>6] <- 6
R> silc.16$hsize6 <- factor(aa, ordered=TRUE)

R> # recode personal economic status
R> aa <- as.numeric(silc.16$pl030)
```

```
R> aa[aa<3] <- 1
R> aa[aa>1] <- 2
R> silc.16$work <- factor(aa, levels=1:2,
R+                        labels=c("working","not working"))

R> # categorize personal net income
R> silc.16$c.netI <- cut(silc.16$net/1000,
R+                       breaks=c(-6,0,5,10,15,20,25,30,40,50,200))
```

In order to reproduce the basic SM framework, the data set `silc.16` is split randomly in two data sets: `rec.A` consisting of $n_A = 4000$ observations and `don.B` with the remaining $n_B = 5522$ units. The data sets `rec.A` and `don.B` share the variables `X.vars`; the person's economic status `y.var`) is available only in `rec.A` while the net income (`z.var`) is available just in `don.B`.

```
R> # simulates a SM framework
R> n <- nrow(silc.16)
R> set.seed(123456)
R> obs.A <- sample(n, 4000, replace=F)
R> X.vars <- c("hsize","hsize6","db040","age","c.age",
R+             "rb090","pb220a", "rb050")
R> y.var <- c("pl030","work")
R> z.var <- c("netIncome","c.netI")

R> # split silc.16
R> rec.A <- silc.16[obs.A, c(X.vars, y.var)]
R> don.B <- silc.16[-obs.A, c(X.vars, z.var)]

R> #rescale weights
R> N <- round(sum(silc.16$rb050)) # est. size of pop(age>15)
R> rec.A$wwA <- rec.A$rb050/sum(rec.A$rb050)*N  # new rough weights
R> don.B$wwB <- don.B$rb050/sum(don.B$rb050)*N  # new rough weights
```

For examples of SM using auxiliary information represented by a third data set containing all the interest variable, a small sample *C* consisting of $n_C = 200$ observations it is selected from the data set `silc.16`.

```
R> # generating artificial sample C
R> set.seed(43210)
R> obs.C <- sample(nrow(silc.16), 200, replace=F)
R> #
R> X.vars <- c("hsize","hsize6","db040","age","c.age",
R+             "rb090","pb220a", "rb050")
R> y.var <- c("pl030","work")
R> z.var <- c("netIncome","c.netI")
R> #
R> aux.C <- silc.16[obs.C, c(X.vars, y.var, z.var)]
R> # rough weights
R> aux.C$wwC <- aux.C$rb050/sum(aux.C$rb050)*round(sum(silc.16$rb050))
R> svy.aux.C <- svydesign(~1, weights=~wwC, data=aux.C)
```

References

Agresti, A (2002) *Categorical Data Analysis, 2nd Edition.* Wiley, Chichester.

Alfons, A and Kraft, S (2012) "simPopulation: Simulation of synthetic populations for surveys based on sample data", R package version 0.4.0 http://CRAN.R-project.org/package=simPopulation

Andridge, RR and Little, RJA (2009) "The Use of Sample Weights in Hot Deck Imputation", *Journal of Official Statistics*, **25**, pp. 21-36.

Andridge, RR and Little, RJA (2010) "A Review of Hot Deck Imputation for Survey Nonresponse", *International Statistical Review*, **78**, pp. 40-64.

Berkelaar, M and others (2013) "lpSolve: Interface to LpSolve v. 5.5 to solve linear/integer programs", R package version 5.6.7 http://CRAN.R-project.org/package=lpSolve

Breiman L., Friedman J. H., Olshen R. A., and Stone, C. J. (1984) *Classification and Regression Trees*. Wadsworth.

Breiman, L. (2001) "Random Forests", *Machine Learning*, **45**, pp. 5-32.

Chen, J and Sitter, RR (1999) "A Pseudo Empirical Likelihood Approach To The Effective Use Of Auxiliary Information In Complex Surveys". *Statistica Sinica*, **9**, pp. 385-406.

Cohen, ML (1991) "Statistical Matching and Microsimulation Models", in Citro and Hanushek (eds.) *Improving Information for Social Policy Decisions – The Uses of Microsimulation Modeling. Vol II. Technical papers*. Commission on Behavioral and Social Sciences and Education (CBASSE), Washington D.C.

Conti, PL and Marella, D and Scanu, M (2012) "Uncertainty Analysis in Statistical Matching", *Journal of Official Statistics*, **28**, pp. 69-88.

D'Orazio, M (2008) "Evaluation of the Accuracy of Statistical Matching", in Eurostat *Report of WP1: State of the Art on Statistical Methodologies for Integration of Surveys and Administrative Data*, ESSnet Statistical Methodology Project on Integration of Survey and Administrative Data, pp. 37-39. http://cenex-isad.istat.it

D'Orazio, M (2011a) "Statistical matching when dealing with data from complex survey sampling", in Eurostat, *Report of WP1. State of the Art on Statistical Methodologies for Data Integration*, ESSnet project on Data Integration, pp. 33-37. http://www.essnet-portal.eu/sites/default/files/131/FinalReport_WP1.pdf

D'Orazio, M (2011b) "Statistical Matching and Imputation of Survey Data with the Package StatMatch for the R Environment" R package vignette http://www.cros-portal.eu/sites/default/files//Statistical_Matching_with_StatMatch.pdf

D'Orazio, M (2012) "StatMatch: Statistical Matching", R package version 1.2.0 http://CRAN.R-project.org/package=StatMatch

D'Orazio, M and Di Zio, M and Scanu, M (2005) "A comparison among different estimators of regression parameters on statistically matched files trough an extensive simulation study", *Contributi Istat*, 10 (2005)

D'Orazio, M and Di Zio, M and Scanu, M (2006a), "Statistical Matching for Categorical Data: Displaying Uncertainty and Using Logical Constraints", *Journal of Official Statistics*, **22**, pp. 137-157.

D'Orazio, M and Di Zio, M and Scanu, M (2006b) *Statistical Matching: Theory and Practice*. Wiley, Chichester

D'Orazio, M and Di Zio, M and Scanu, M (2008) "The Statistical Matching Workflow", In Eurostat *Report of WP1: State of the Art on Statistical Methodologies for Integration of Surveys and Administrative Data*, ESSnet Statistical Methodology Project on Integration of Survey and Administrative Data, pp. 25-26. http://cenex-isad.istat.it

D'Orazio, M and Di Zio, M and Scanu, M (2010) "Old and new approaches in statistical matching when samples are drawn with complex survey designs", *Proceedings of the 45th Riunione Scientifica della Societá Italiana di Statistica*, 16–18 June 2010, Padova, Italy

D'Orazio, M and Di Zio, M and Scanu, M, (2012) "Statistical Matching of Data from Complex Sample Surveys", European Conference on Quality in Official Statistics - Q2012, 29 May- 1 June 2012, Athens, Greece

Efron B., Hastie T., Johnstone I., and Tibshirani R. (2004) "Least angle regression", *Annals of Statistics*, Volume 32, Number 2 (2004), pp. 407-499.

Genz, A, Bretz, F, Miwa,T, Mi X, Leisch, F, Scheipl, F, and Hothorn T (2013). mvtnorm: Multivariate Normal and t Distributions. R package version 0.9-9995. http://CRAN.R-project.org/package=mvtnorm

Gertheiss J. (2011) "Testing linearity and relevance of ordinal predictors", *Electronic Journal of Statistics*, Vol. 5 (2011), pp. 1935–1959

Gower, JC (1971) "A general coefficient of similarity and some of its properties", *Biometrics*, **27**, pp. 623-637

Groll, A. and G. Tutz (2011). "Variable selection for generalized linear mixed models by L1-penalized estimation". *Technical Report* 108, Ludwig-Maximilians-University

Harrell F.E. (2001) *Regression Modeling Strategies with Applications to Linear Models, Logistic Regression, and Survival Analysis.* Springer-Verlag, New York.

Harrell, FE and others (2013) "Hmisc: Harrell Miscellaneous", R package version 3.10-1 http://CRAN.R-project.org/package=Hmisc

Hornik, K (2005) "A CLUE for CLUster Ensembles", *Journal of Statistical Software*, **14/12**, http://www.jstatsoft.org/v14/i12/

Hornik, K (2013) "clue: Cluster ensembles", R package version 0.3-46, http://CRAN.R-project.org/package=clue

Korn, EL and Graubard, BI (1991) *Analysis of Health Surveys*. Wiley, New York.

Kovar, JG and MacMillan, J and Whitridge, P (1988) "Overview and strategy for the Generalized Edit and Imputation System", Statistics Canada, *Methodology Working Paper*, No. BSMD 88-007 E/F

Little R.J.A., Rubin D.B. (2002) *Statistical Analysis with Missing Data, 2nd Edition*. Wiley, New York.

Lumley, T (2012) "survey: analysis of complex survey samples", R package version 3.28-2. http://CRAN.R-project.org/package=survey

Meyer, D and Buchta, C (2013) "proxy: Distance and Similarity Measures", R package version 0.4-10. http://CRAN.R-project.org/package=proxy

Moriarity, C and Scheuren, F (2001) "Statistical Matching: a Paradigm for Assessing the Uncertainty in the Procedure", *Journal of Official Statistics*, **17,** pp. 407-422

Moriarity, C and Scheuren, F (2003) "A note on Rubin's statistical matching using file concatenation with adjusted weights and multiple imputation", *Journal of Business and Economic Statistics*, **21**, pp. 65-73

R Core Team (2013) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing Vienna. http://www.R-project.org/

Rässler, S, (2002) *Statistical Matching: a Frequentist Theory, Practical Applications and Alternative Bayesian Approaches*. Springer-Verlag, New York.

Rässler, S, (2003) "A Non-iterative Bayesian Approach to Statistical Matching". *Statistica Neerlandica*, **57**, pp. 58-74.

Renssen, RH (1988) "Use of statistical matching techniques in calibration estimation", *Survey Methodology*, **24**, pp. 171-183.

Rubin, DB (1986) "Statistical matching using file concatenation with adjusted weights and multiple imputations", *Journal of Business and Economic Statistics*, **4**, pp. 87-94

Sarle W.S. (1990) "The VARCLUS Procedure". *SAS/STAT User's Guide, 4th Edition*. Cary NC: SAS Institute, Inc.le

Särndal, CE and Swensson, B and Wretman, J (1992) *Model Assisted Survey Sampling*. Springer-Verlag, New York.

Särndal, CE and Lundström, S (2005) *Estimation in Surveys with Nonresponse*. Wiley, Chichester.

Scanu, M (2008) "The practical aspects to be considered for statistical matching", In Eurostat *Report of WP2: Recommendations on the use of methodologies for the integration of surveys and administrative data*, ESSnet Statistical Methodology Project on Integration of Survey and Administrative Data, pp. 34-35. http://cenex-isad.istat.it

Singh, AC and Mantel, H and Kinack, M and Rowe, G (1993) "Statistical matching: use of auxiliary information as an alternative to the conditional independence assumption", *Survey Methodology*, **19**, pp. 59-79.

Wu, C (2004) "Combining information from multiple surveys through the empirical likelihood method", *The Canadian Journal of Statistics*, **32**, pp. 15-26